Fall 2013

Estimation of Variation For High-throughput Molecular Biological Experiments With Small Sample Size

Danni Yu
Purdue University

Graduate School ETD Form 9
(Revised 12/07)

PURDUE UNIVERSITY
GRADUATE SCHOOL
Thesis/Dissertation Acceptance

This is to certify that the thesis/dissertation prepared

By Danni Yu

Entitled
Estimation of variation for high-throughput molecular biological experiments with small sample size

For the degree of Doctor of Philosophy

Is approved by the final examining committee:

Olga Vitek
Chair

Bruce Cragi

Hao Zhang

Michael Yu Zhu

To the best of my knowledge and as understood by the student in the *Research Integrity and Copyright Disclaimer (Graduate School Form 20)*, this thesis/dissertation adheres to the provisions of Purdue University's "Policy on Integrity in Research" and the use of copyrighted material.

Approved by Major Professor(s): Olga Vitek

Approved by: Jun Xie 08/29/2013
Head of the Graduate Program Date

ESTIMATION OF VARIATION FOR HIGH-THROUGHPUT MOLECULAR BIOLOGICAL EXPERIMENTS WITH SMALL SAMPLE SIZE

A Dissertation

Submitted to the Faculty

of

Purdue University

by

Danni Yu

In Partial Fulfillment of the

Requirements for the Degree

of

Doctor of Philosophy

December 2013

Purdue University

West Lafayette, Indiana

I dedicate this dissertation to my warm and lovely family. My husband, Luo Si, has been so kindly providing me the strongest support in these many years of research. My parents and my in-laws, especially my mother Xiaozhong Han, have been giving me the best help with their unlimited effort and heart for baby-sitting and housework. My two little boys bring me the sweetest happiness in my life. My grandmother Huizhong Li, uncle Sherwin Han, and auntie Stephanie Han, have wished and helped the best for my career.

ACKNOWLEDGMENTS

I own a debt of gratitude to my advisor, Dr. Olga Vitek, for the insightful guidance, financial support, invaluable advice and strong encouragement that have been given to me in these six years. Her willingness to spend time with me so generously has been very much appreciated. Being not only a wonderful advisor but also a great mentor, Dr. Vitek taught me how to approach research problems with scientific attitude, how to communicate with scientists in multi-disciplinary areas, how to write conference and journal papers, and how to share success with collaborators. This research experience under Dr. Vitek's tremendous help is extremely valuable to me and becomes a treasure that will be saved in the rest of my life.

Special and sincere thanks to the other members of my committee members, Dr. Bruce Craig, Dr. Hao Zhang, and Dr. Michael Yu Zhu for their very helpful comments and suggestions on my thesis work. I hope to further express my gratitude to Dr. Zhu for the valuable information, suggestions and help on my Ph.D study and research that he has been generously giving to me. Without the contribution from my committee members, I will never be able to complete this dissertation.

I hope to thank Dr. David E. Salt for his advice in Biological science and three-years financial support, and Dr. Laura P. Sands for her guidance in Quantitative Psychology and two-years financial support. Special thanks to Dr. Wolfgang Huber for the very kind guidance, suggestions and financial support.

Many thanks to the department of Statistics which provides a very kind academic environments for students to quickly grow up. I hope to especially thank Dr. Rebecca W. Doerge because she has been providing me so much help and support that are essential for me to conquer difficulties. I am also very grateful of Dr. Mary Ellen Bock, Dr. Herman Rubin, Dr. Chuanhai Liu, Dr. Dabao Zhang, Dr. Huiping Xu, Dr. John Danku, Dr. Ivan Baxter, etc., for their very kind and generous help on my Ph.D study. Many thanks to Dr.

Maurice Burg, Dr. Joan Ferraris, Dr. Yuichiro Izumi, Taruna Singh, Dr. Jinxi Li, Chester Williams, Dr. Xiaoming Zhou, Dr. Zheng Li and Dr. Jun Zhu for the very kind help starting from the time when I served an internship at National Institute Health. Thanks Dr. Wolfgang Huber, Dr. Simon Anders, Dr. Joseph Barry, Dr. Bernd Fischer and all the other colleagues when I visited Dr. Huber's lab at European Molecular Laboratories. I also hope to thank Dr. Anne-Claude Gingras, Dr. Brett Larsen, Dr. Steve Tate, Dr. Ludovic Gillet, Hannes Röst for their very kind help and collaboration with me in the research of protein quantification.

Finally, I hope to thank Dr. Douglas Crabill and his team for the professional help on IT solutions, and thank Ms. Marian Duncan, Ms. Diane Martin, Ms. Becca Pillion, Ms. Shaun Ponder and Ms. Mary Roe for the very kind administrative support.

TABLE OF CONTENTS

Page

Page

LIST OF TABLES

Table Page

LIST OF FIGURES

Figure Page

Figure Page

ABBREVIATIONS

MLE	maximum likelihood estimator
MM	method of moments
DIA	Data-Independence Acquisition
DNA	deoxyribonucleic acid
ICP-MS	Inductively Coupled Plasma spectroscopy combined with Mass Spectroscopy
IDA	Information-Dependent Acquisition
MRM	Multiple Reaction Monitoring for multiple SRM transitions
MS	Mass Spectrometry
MS/MS	tandem mass spectrometry
MS1	precursor mass spectra for peptides
MS2	tandem mass spectra for fragments of peptides
m/z	mass-to-charge ratio
NGS	Next-Generation Sequencing
RNA-seq	whole transcriptome shotgun sequencing
SRM	Selected Reaction Monitoring for quantitative proteomics
SWATH	Sequential Window Acquisition of All THeoretical fragment-ion spectra
XIC	eXtracted Ion Chromatogram

ABSTRACT

Yu, Danni Ph.D., Purdue University, December 2013. Estimation of Variation For High-throughput Molecular Biological Experiments With Small Sample Size. Major Professor: Olga Vitek.

Motivation: In the quantification of molecular components, a large variation can affect and even potentially mislead the biological conclusions. Meanwhile, the high-throughput experiments often involve a small number of samples due to the limitation of cost and time. In such cases, the stochastic information may dominate the outcome of an experiment because there may not be enough samples to present the true biological information. It is challenging to distinguish the changes in phenotype from the stochastic variation.

Methods: Since the biological molecules have been quantified with different technologies, different statistical methods are required. Focusing on three types of important high-throughput experiments, this thesis proposes novel solutions to reduce noise and increase the accuracy of molecular discovery.

i) In the large-scale perturbation screens, thousands of mutant strains on hundreds of plates are separately profiled in hundreds of days (or batches). For each mutant strain, only a small number of samples are profiled. The artificial noise mainly consists of additive and multiplicative effects due to plates and batches. We propose a linear mixed-effect modeling framework based on experimental designs with at least two control samples. These are involved in a normalization and variance estimation procedure for the purpose of reducing the noise from data and scoring the true biological phenotype.

ii) In the RNA-seq experiments, fragments of greater than thousands of genes in 4$\sim$8 samples on a flow cell can be sequenced in one day. The additive and no-additive effects due to the large number of plates do not typically present in the data. The gene-wise variance between samples consists of both the expectation and dispersion of gene counts. Due to

stochastic noise, some of gene-wise dispersion are under or over estimated. This may lead to misinterpretation of the biological phenotype. We propose a shrinkage estimator of dispersion under Negative Binomial models to regularize the estimates towards a value calculated from common information across genes.

Lastly *iii)* in the MS/MS experiments with SWATH acquisition, more than 10 thousand spectra in a run can be sequentially obtained in about 120 minutes. The summed up intensity across all the signals within a tiny m/z bin is used to identify fragments of each peptide. As a result, the interference noise within the m/z bins leaves undetected and misleading ambiguity in protein quantification. The solutions previously proposed for perturbation screens and RNA-seq experiments can not be used for SAWTH acquisition because the property of the data is different. In order to remedy such defects, a new approach is proposed to quantify the homogeneity (opposite to interference) among the co-elusion traces of molecules within the m/z bins. Since correct signals of a fragment share a homogeneous peak shape, we propose to utilize the p-value of one-side test on the second order coefficient in a linear quadratic model. The coefficient accounts for the curved shape in a linear regression procedure. The p-value represents the strength of concave pattern across those peaks of a fragment.

Results: The evaluation results of different experiments with each of the three technologies illustrate that the proposed solutions outperform several existing methods.

1. INTRODUCTION

Genes and proteins are the fundamental molecules for living organisms. Biological information saved by millions of DNAs on genes are transferred and utilized to produce proteins [1]. Fulfilling their various functions leads to abundant changes of those molecules. Consequently, the quantification studies of genes, proteins and the products of gene perturbations facilitate the discovery of functional genes in life science. However, the scored phenotype consists of not only the biological changes but also sources of variation that substantially affect the biological conclusions.

Since different techniques and experimental methods have been developed to account for the various characteristics of genes and their products, it is challenging but required for statisticians to accurately estimate and reduce the noise from data using different statistical methods.

This thesis focuses on estimating the variation of quantitative phenotype collected in three types of important high-throughput experiments. The souse of major variations are carefully studied. We propose different and appropriate statistical solutions to reduce noise in such experiments. Problems and contributions of these projects are introduced in Chapter 1.1−Chapter 1.3. Two of these projects and the extended work have been published [2–4].

1.1 Genome-wide perturbation screens

1.1.1 Statement of biological problem

A main purpose of studying the perturbation screens in this project is to identify the changes in the elements' abundance associated with the lost-of-function or over expression of a gene. In the study of genome-wide perturbation screens, the phenotypes of thousands

of such mutant samples are quantified under various conditions. Consequently, those scored phenotypes are subject to both biological and technical variation.

In these experiments, it is practical to include a small number of replicates for each mutant, such as $n = 4,\ 8,\ 16$. Randomizing the order of genome-wide mutant samples throughout the screens removes the systematic pattern between plates. This is because similar phenotypes may be obtained in the mutant samples when perturbing the genes close to each other on a chromosome. To implement the experiment, an available instrument only processes three 96-well plates in a day. Those mutant samples have to be distributed into hundreds (e.g. 300 480) of plates and profiled in hundreds (e.g. 100 160 days) of days (i.e. batches). It produces shifts and alters the scales of phenotype in different plates and batches. These additive and non-additive random noise due to the large number of batches and plates nested in batches must be addressed for accurate detection of hits.

1.1.2 Statement of statistical problem

A practical and implementable experimental design for the high-throughput perturbations screens is required. The experiments should not only collect the biological information in mutant samples but also facilitate the estimation of the additive and non-additive variation between plates and batches.

For the purpose of reducing the noise, a statistical model is required to account for the random effects in batches and in plates nested in batches. In addition to the additive noise, the statistical mutant $\times$ batch interaction and the statistical mutant $\times$ plate interaction should be used to account for the non-additive or multiplicative variation. However, the estimation of those parameters can not be directly obtained because currently it is impossible to collect technical replicate measurements of each mutant sample in every plate.

1.1.3 Contributions

To solve the above-mentioned problems, the following procedure is proposed in [2].

- We propose an experimental design that involves at least two negative quadruplicate control samples. These control samples are repeated in all of the plates. For the purpose of evaluation, we also recommend including the other quadruplicate control samples and repeat them in all the plates. For a genome-wide perturbation screen, mutant samples should be quadruplicated and nested in plates.

- A linear mixed-effect modeling fitted in the first negative control samples is proposed to estimate the additive effects.

- Another linear mixed-effect modeling fitted in the second negative control samples is proposed to estimate the non-additive effects.

- A score is proposed to summarize the biological phenotype of each mutant sample. The score is based on the normalization after using the first model and the variance estimation using the second model.

The extended work of obtaining biological conclusion based on the results in [2] is published in [3]. The experimental datasets are published at www.ionomicshub.org. The open-source R package `HTSmix` is available at http://www.stat.purdue.edu/∼ovitek/HTSmix.html.

1.2 RNA-seq experiments

1.2.1 Statement of biological problem

The digital counts of reads for gene expression are produced in RNA-seq experiments. In one day, genome information can be obtained in a small number of samples nested in one or several flow cells. The noise due to large number of plates and batches are not the major concern in this project when the total number of samples are small. Instead, the stochastic variation in genes and samples is the major focus.

1.2.2 Statement of statistical problem

The gene abundance is quantified in counts. The traditional statistical methods for gene expression analysis with microarrays may not be directly applied since the gene abundance was based on a continuos variable.

Currently, the counts are frequently modeled by the Negative Binomial distribution. The model is utilized to distinguish the systematic changes in gene expression between conditions from noise. However, the per-gene estimates of the dispersion parameter is often not reliable in the experiments with small sample size.

1.2.3 Contributions

The aim of our research is to provide a simple but effective approach for characterizing the variation in the counts of reads, we propose in [4] to

- calculate the initial estimates for per gene dispersion using the method of moments,
- shrunk all the estimates towards the common information borrowed across genes,
- estimate such common information by the value within the range of initial estimates that minimizes the average squared difference between the initial estimates and the shrinkage estimates.

Without requiring extra modeling assumptions, the proposed method is computationally efficient, and compatible with the exact test of differential expression. The open-source R package `sSeq` is available at http://www.stat.purdue.edu/∼dyu/sSeq and Bioconductor.

1.3 MS/MS with SWATH aquisition

1.3.1 Statement of biological problem

The study of changes in protein abundance relies on the identification and quantification of the peptides (i.e. fragments of proteins) at MS1 level or the fragments of peptides at

MS2 level. In this project, we focus on the data collected in MS/MS with SWATH acquisition, which is a novel technique based on data-independent approach. Given a spectral library including the prior known targets, the fragments of peptides can be identified and quantified based on the MS2 spectra. This method produces the results with reasonably good sensitivity and specificity for protein/peptide identification [5].

However, in complex experiments with a large number of proteins, misleading information can be produced by noise interference in the fragments that are identified as a peptide, but actually from a different source. Such fragments often co-elute and share similar mass-to-charge ratios (m/z). To distinguish the true information from the interference noise is challenging.

1.3.2 Statement of statistical problem

The existing method such as openSWATH identifies the peptides based on the extracted-ion chromatograms (XICs). Specifically in this project, a two-dimensional bin spans around 5 minutes in retention time and 50ppm in spectra. Counts of all the molecules observed in the 50ppm window are summed up within each time point. Finally, the XIC of an identified fragment consists of the total counts at each time point and a small range of retention time. This existing method helps identify the peptides but leaves potential ambiguity into the fragments' quantification. This is because molecules within the 50ppm belonged to different fragments can also produce high intensity in a XIC and encourage the plausible identification.

Consistent peak pattern among those molecules is a key concept of judging a true fragment. However, it is challenging to statistically model the shared pattern between those molecules.

1.3.3 Contributions

To increase the accuracy of quantification within fragment and account for the interference noise between mis-identified fragments, we propose to

- separate the XIC of a fragment into miniXICs for the molecules within the 50ppm window,
- model the association between retention time and the molecules' intensity with one quadratic regression,
- weigh the strength of peak shape pattern among the molecules using the p-value of testing the second-order coefficient in the quadratic regression model, and then
- extend the quadratic regression by adding a fixed-effect factor (i.e. fragments), and fit the model in the data including the miniXICs of all the identified fragments of a peptide.

Evaluation results upon real experiments illustrate the quantification combined with the proposed method empirically improves the sensitivity and specificity in comparative study of detecting changes in protein abundance.

1.4 Outline

This dissertation is organized as follows. Chapter 2 describes the experiments of perturbation screens, the related work and the proposed solution for noise reduction and variance estimation in the quantification of element abundance. Chapter 3 presents the problems of gene quantification in RNA-seq experiments and the proposed approach of regularizing the dispersion estimates. Chapter 4 describes the data-independent acquisition experiments with SWATH technology for protein quantification and proposes a statistical solution to score the strength of concave pattern across traces of signals with fragments. Finally, the results are summarized and future work is discussed in Chapter 5.

2. GENOME-WIDE PERTURBATION SCREENS

2.1 Introduction

[1] In functional biology [6,7], and in biomedical [8] and biopharmaceutical research [9], the technology of perturbation screening [10, 11] has been used to observe a variety of phenotype in model organisms subjected and associated to the stresses. Two major types of perturbation are either external such as heat shock or chemical treatments, or internal such as genetic disruption or deletion of genes. The measurement on the phenotypes can be univariate (e.g. cell growth rate or activity of a reporter gene), low-dimensional (e.g cellular morphology and ionomics), or high-dimensional (e.g. gene expression or protein abundance). Insightful information on the function of living organisms [12, 13] can be obtained when the perturbation study is performed in a genome-wide scale.

However, a primary concern in the investigations of genome-wide screens is how to accurately quantify the sources of variation in these high-throughput data, and increase the reproducibility between experiments. Because measuring the phenotypes of one-by-one genetic perturbations takes several months when working with Inductively Coupled Plasma Mass Spectrometry (ICP-MS), the fundamental principles of statistical methods can not be directly utilized. Especially, there are typically very small number of replicates (n=1, 2, 3, 4) incorporated in perturbation screens due to the limitation of instruments. Fully randomizing the order or the replicates in genome-wide scale is impossible. Furthermore, the problem compounded with different experimental characteristics such as instruments, labor, and reagents can also not be removed from large-scale screens. Consequently, the sources of variation existing in the measurements include both the natural between-sample variation and the technical variation for handling samples and procedures. It is an essential and non-trivial step to interpret and account for the specifics in such kind

[1] Copy rights are released with Publisher's permission.

of high-throughput perturbation screens. It is particularly challenging in such situations to distinguish the systematic signal from noise.

In this chapter, we focus on the problem in cases of low-dimensional phenotypes which are sensitively affected in an non-ignorable proportion of the samples. We propose a framework with statistical modeling [2] to reduce the noise and accurately interpret the phenotypes in high-throughput screens that have a limited number of replicates.

2.2 Background

The five steps involved in the study with perturbation screens are experimental design, normalization, phenotype summarization for each sample over multiple replicates with estimation of the associated variation, determination of “hits", i.e. samples with systematic changes in phenotype, and evaluation. Details about these steps are provided in the following paragraphs.

2.2.1 Experimental design

We overviewed a typical design of a perturbation screen in Fig. 2.1. Among the types of microtiter plates (e.g. 6, 24, 96, 384, or 1536 sample wells), the 96-well plates are typically used by the ICP-MS instrument to balance the work load of a labor and the control of environmental effects. In order to obtain high throughput, only a small number of replicates for each mutant samples can be profiled, such as four replicates per sample nested in a plate (recommended by [14]). To remove the variation due to the interaction between sample and plate, all the replicates of a sample are systematically allocated to the same plate. To enable genome-wide perturbation screening, hundreds or even thousands of plates are required. Therefore, due to the limitation of biological material and capacity of equipment, it is necessary to handle those large numbers of plates in batches.

In [15], it is observed that the batches-and-plates effect can systematically distort the scored phenotypes. The variation in rows and columns on a plate [15] or the excessive evaporation of media around the edges [16] produce the within-plate effects. These artifacts

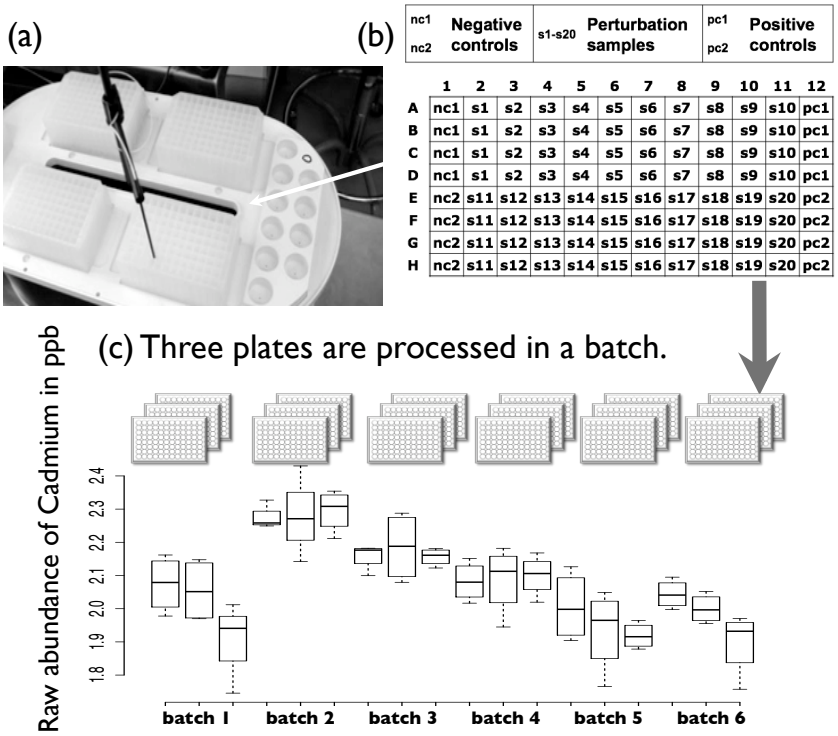


Fig. 2.1.: Experimental design of a perturbation screen.

(a) An instrument processes three plates in a day. The plate has 96 wells in which the abundance of 14 or 17 elements were measured one-by-one in the order randomized by the instrument itself. (b) Two quadruplicated negative controls were stored in the first column of each plate, and the other two quadruplicated positive controls were located in the last column. The remaining columns were used for twenty quadruplicated mutant samples with genetic perturbations. The control samples were repeated in all the plates, but the mutant samples were different in different plates. (c) An example uses the boxplot for the first 18 plates of the screen to visualize the distribution of the scored abundance of one element (i.e. Cadmium) for a positive control named as YPR065W (or ROX1). Three plates were processed in a day termed as a batch. It clearly indicates that there are systematic effects of plates and batches on the phenotype.

can be accounted for using one or more control samples included in all plates. Those control samples should be negative, i.e. the samples are not perturbed and naturally survive in the wild environment, or positive, i.e. the samples are known to have changes in the phenotype. In order to limit the artificial effects of evaporation on the perturbed mutant samples, it was recommended [15] to allocate the controls samples around the edges of each plate.

In practical study with high-throughput perturbation screens, due to the limited capacity of plates, the absence of between-plate replication, and the small number of within-plate replication of perturbed samples and control samples, it is challenging to eliminate the experimental artifacts for accurate quantification on phenotype. In this chapter Chapter 3.3, we argue that several existing statistical methods have not well used the control samples in these situations. By separately using two or more distinct positive or negative controls, we are able to obtain a more specific detection of hits.

2.2.2 Normalization

After samples are processed by an instrument, the outlying measures or died samples are eliminated for quality control. In order to have the measured phenotype comparable across samples, batches and plates, a normalization procedure is required to account for confounding and experimental artifacts.

Sample-based and control-based are two approaches of normalization that are frequently used. By assuming that the majority of perturbations do not affect the phenotype, sample-based normalization methods normalize to a pool of samples. The examples such as B-score [17], Z-score, and plate-wise median [18] are shown as follows. Those methods were reviewed in [15] and shown as follows.

B score specifies a linear model when the additive effects of rows and columns, $R_{i,p}$ and $C_{j,p}$, within plate p exist in model Eq. (2.1),

$$X_{ij,gkp} = \mu_p + R_{i,p} + C_{j,p} + \varepsilon_{ij,gkp}, \tag{2.1}$$

where the average phenotype within plate p is denoted by μ_p, $X_{ij,gkp}$ is the intensity of replicate k of mutant g on row i and column j in plate p, and the non-systematic error $\varepsilon_{gkp} \sim \mathcal{N}(0,\ \sigma^2_{gkp})$. Separately for each plate, all the measurements within a plate are used to estimate the parameters in Eq. (2.1). Using a robust method for sample averages (i.e. Tukey median polish), the residue between the scored phenotype and the estimated expectation is obtained in Eq. (2.2). Finally, a B score for the kth replicate of mutant g in plate p is the model residuals in the unit of its median absolute deviation.

$$r_{gkp} = X_{gkp} - [\hat{\mu}_p + \hat{R}_{i,p} + \hat{C}_{j,p}]. \tag{2.2}$$

$$\text{Bscore}_{gkp} = \frac{r_{gkp}}{\text{median}\left(\ |r_{gkp} - \text{median}_{(r_{gkp})_p}|\ \right)_p}. \tag{2.3}$$

Z score is the difference between the scored phenotype and the average of all samples within a plate, which is then scaled in the unit of plate-specific standard deviation. In Eq. (2.4), $\bar{X}_{..p}$ and $s_{..p}$ are the mean and the standard deviation of all the samples within plate p.

$$\text{Zscore}_{gkp} = \frac{X_{gkp} - \bar{X}_{..p}}{s_{..p}} \tag{2.4}$$

Plate-wise median normalization (pmNorm) is just the scored phenotype in the unit of its within-plate median, and formularized as

$$\text{pmNorm}_{gkp} = \frac{X_{gkp}}{\text{median}(X_{gkp})_p}. \tag{2.5}$$

Quantile normalization is a method popularly utilized for gene expression in microarrays experiments [19,20]. The method is extended to perturbation screens [21] for normalization between plates. It assumes that phenotypes between plates share the same empirical distribution G. The observed phenotype X_{gkp} is firstly transformed into G and then reversely tranformed back through F which is the averaged distribution of sample quantiles across all plates.

$$r_{gkp} = F^{-1}\left(G(X_{gkp})\right) \tag{2.6}$$

The other sample-based normalization methods can be utilized to account for the batch effect, such as principle analysis [22], and surrogate variable analysis [23].

Since sample-based normalization uses the entire collection of samples in the perturbation experiment, it maximizes the degree of freedom for parameter estimation and produces an accurate output of the normalized phenotype. However, in the situation when a large number of perturbations affect the phenotype and especially when the secondary and confirmatory screens are performed, the assumption of sample-based normalization is not appropriate. As an alternative, control-based normalization should be used [24]. Description on several existing control-based normalization is provided here.

Normalized percent inhibition (NPI) quantifies the inhibition with the difference between mean of positive controls' phenotype $\bar{c}_p^+$ and a sample's scored phenotype X_{gkp} in the unit of difference between mean of positive controls and mean of negative controls $\bar{c}_p^-$. All measurements were normalized within each plate.

$$\text{NPI}_{gkp} = \frac{\bar{c}_p^+ - X_{gkp}}{\bar{c}_p^+ - \bar{c}_p^-}, \tag{2.7}$$

Percent of control mean (pocMean) in Eq. (2.8) normalizes sample phenotype X_{gkp} to mean of positive controls $\bar{c}_p^+$ by considering the measurements of positive controls as a baseline in each plate. **Percent of control median (pocMed)** formularized in Eq. (2.9) normalizes sample phenotype to median of negative controls $\tilde{c}_p^-$ with the information that those controls do not affect the phenotype. Since negative controls are often affected by outliers, median of measured phenotype across controls is used.

$$\text{pocMean}_{gkp} = \frac{X_{gkp}}{\bar{c}_p^+} \times 100, \tag{2.8}$$

$$\text{pocMed}_{gkp} = \frac{X_{gkp}}{\tilde{c}_p^-} \times 100, \tag{2.9}$$

In perturbation screen, each plate only can include a small number of controls. The existing normalization method based on controls only account for limited types of experimental artifacts. As a result, highly variable estimates of bias can not be eliminated. The performance of seven sample-based and control-based normalization methods were compared in [16]. As a conclusion in the words of [24], "no single method excelled" in all situations. The softwares, such as the open-source Bioconductor packages `RNAither` [25]

and `cellHTS2` [26], offer the implementation for several above-mentioned normalization methods.

In this chapter, it is demonstrated that, compare to several existing methods, we are able to improve the accuracy of results using control-based normalization for the experiments in which a large proportion of samples with changed phenotypes are shown in screens. We argue that at least two controls should be involved, one used for normalization and another one used for estimation of between-plate variation.

2.2.3 Summarization of phenotypes, and estimation of variation

In order to borrow information across all the samples in genome-wide perturbation study, a comparable summarizing score over the replicates for a biological sample needs to be provided such as an averaging value. Meanwhile, it is also important to estimate the associated variation, which enables us to distinguish random variation from perturbation-related changes in the phenotype. Sample variance or its robust alternatives are the estimator that most existing methods have used [18]. The other method, such as Empirical Bayes approach for variance stabilizzation recommended in [15], is directly applicable to the context of perturbation screens. The approach was originally introduced in the study of gene expression with microarray experiments [27, 28].

However, when there are not between-plate replicates in genome-wide study, the above-mentioned methods for estimation of variation can have serious deficiencies affected by the unknown and potentially sensitive phenotypes. In such kind of experiments that allocate all replicates of a biological sample in the same plate, only the estimates of within-plate variation can be provided. Those methods actually assume that the variation in a plate is enough to account for the entire between-plate variation in the normalized phenotypes.

In Chapter 2.3.3, we demonstrate that this assumption over-simplifies the structure of variation in high-throughput perturbation screens. and this problem is rarely verified in previous literature. We also note that estimates of plate-and batch-specific bias are subject to uncertainty when normalizing quantities are estimated from a small number of obser-

vations. Moreover, in different biological samples, the plate-and-batch-specific effect on the phenotype can be different. It further contributes to the variation. We illustrate that it is important for accurate determination of hist to account for this residual variation in an appropriate way.

2.2.4 Determination of hits

The certain value μ_0 that is compared in a null hypothesis $H_o : \mu_g = \mu_0$ is typically the phenotype of a control or the average phenotype of all stressed samples, where μ_g is the average phenotype over replicates of one perturbation sample. A test statistic, such as the Student T or the moderated T, is utilized to compare the summary quantification of the phenotype with its estimate of variation. Depending on the experiment, the reference distribution of the statistic is assumed as Student or Normal, or is estimated empirically based on controls. Non-parametric alternatives , e.g. the Mann-Whitney test and the Rank Product test [25] can also be used, but the power is lower.

Student T statistic is based on the summarization by sample average, and on variance estimation by sample variance. The test statistic is defined as

$$T_g = \frac{\bar{v}_{g\cdot\cdot} - c}{\sqrt{s_g^2/n_g}}, \text{ where } s_g^2 = \frac{1}{n_g - 1}\sum_{k=1}^{n_g}(v_{gkp} - \bar{v}_{g\cdot\cdot}) \tag{2.10}$$

where n_g is the number of replicates of mutant g in plate p.

Moderated T statistic was originally proposed in the context of gene expression microarrays [27], and improves upon the estimate of variance s_g^2 for experiments with a small number of replicates. The approach assumes a Scaled Inverse Chi-square prior distribution of σ_g^2 (or, equivalently, an Inverse Gamma distribution), and uses an Empirical Bayes procedure to derive the test statistic. Formally, the approach assumes

$$\frac{1}{\sigma_{\tilde{\varepsilon}_g}^2} \overset{iid}{\sim} \frac{1}{d_0 s_0^2}\chi_{d_0}^2, = Gamma(\frac{d_0}{2}, \frac{d_0 s_0^2}{2}), \tag{2.11}$$

the the moderated T statistic is

$$T_g = \frac{\bar{B}_{g\cdot bp} - c}{\sqrt{\tilde{s}_g^2 / n_g}}, \text{ where } \tilde{s}_g^2 = \frac{(n_g - 1)s_g^2 + d_0 s_0^2}{(n_g - 1) + d_0}. \tag{2.12}$$

s_0^2 and d_0 in the expression above are the degrees of freedom parameter and the scale parameter of the prior distribution, which are estimated empirically from the entire collection of mutants in the screen. In other words, the joint analysis of all the mutants provides an additional information on the variation, and is equivalent to a prior dataset with estimated variance s_0^2 based on d_0 degrees of freedom.

After calculating the test statistics, it is a non-trivial problem to select cutoff for determination of hits and control the rate of false positive hits at the desired level. Methods in multiple testing procedures such as [29] or [30] can be directly used for controlling the False Discovery Rate (FDR). As an alternative, a specialized Bayesian procedure was developed in [31] to directly model the probabilities of phenotypes and controls FDR. In [32], ordered Z scores was designed by a tool called RNAiCut for automated identification of pathway-relevant hits. Although all these approaches are appropriate, the choice of the test statistic and its estimate of variation directly affect their sensitivity and specificity.

2.2.5 Evaluation

Since the true biological information is typically unknown in experimental datasets, evaluating the performance and efficacy of methods is challenging. We take the advantage of multivariate phenotypes in the situation where Pearson correlations of summarized phenotypes between replicate plates for one control measure the reproducibility of biological samples. If a method accurately and appropriately process the perturbation screens, the reproducibility should be high and the evaluation value should be close to 1. Other wise, the value is close to 0 and the method is plausible. Such multivariate phenotypes of both control-based and sample-based are used for evaluation in Chapter 3.5.

2.3 Methods

We focus on high-throughput perturbation screens with one-dimensional or low-dimensional quantitative phenotypes in this chapter. Specifically we illustrate the proposed method with experiments of genetic perturbations, and refer to the screened samples as mutants. However, the discussion is applicable to all perturbation types. The highly disruptive perturbations or sensitive phenotypes, where we cannot expect a small number of hits, are particularly considered.

The proposed method is a stepwise interpretation procedure based on linear mixed-effects models. This stepwise modeling is computationally efficient compare to global mixed-effects modeling that fits the entire dataset. The approach of stepwise procedures was successfully applied in the context of gene expression microarrays [33, 34], and is similarly effective for perturbation screens.

2.3.1 Experimental design

The experiments that we utilize to illustrate the methods were processed in 96-well or similar plates. All replicates of a sample were profiled in the same plate. Suggested in [15], all within-plate allocations of samples, and a small number of biological replicates, e.g. 4 recomended by [14], can be used.

The proposed approach requires the presence of more than one control samples, profiled in all batches and all plates. Normalization of the phenotype across batches and plates uses the first control. Estimate of the associated variation uses the second control to derive the summary statistic for each mutant. It is beneficial for evaluation if one or two additional control samples, complementing the previous two, are also profiled.

2.3.2 Normalization

Basic model-based normalization. A scored univariate phenotype is denoted as X_{gkbp}, where g is the index of the mutant in which a specific *gene* was perturbed, k is the index

of a replicate for that mutant g, b is the *batch* index and p is the *plate* index where the mutant replicate k was located. In the situation that multivariate phenotypes are profiled for each sample, we consider each dimension separately, and use the convention that X_{gkbp} represents one particular dimension.

Technical variation such as batches and plates, and biological variation are the major sources of variation in a screen. In the normalization model we assume that these effects are non-systematic Normal random variables represented in the following linear model

$$X_{gkbp} = \mu_g + B_{gb} + P(B)_{gp} + \varepsilon_{gkbp} \tag{2.13}$$
$$B_{gb} \overset{iid}{\sim} \mathcal{N}(0, \sigma^2_{B_g}), P(B)_{gp} \overset{iid}{\sim} \mathcal{N}(0, \sigma^2_{P_g}),\ \varepsilon_{gkbp} \overset{iid}{\sim} \mathcal{N}(0, \sigma^2_{\varepsilon_g})$$

where X_{gkbp} is the intensity of replicate K of mutant g on plate p nested in batch b, B_{gb} is a random factor due to batch effect, $P(B)_{gp}$ is a random factor due to plate effect nested within batches, and ε_{gkbp} is the random factor due to model error, and it includes the rest of biological and technical variation that the other two random variables do not account for. All the three random variables in Eq. (2.13) , B_{gb}, $P(B)_{gp}$ and ε_{gkbp}, are independent of each other. Particularly, g=1, 2, 3, 4, represent the indexes for the first, second, third and fourth controls, respectively. And g=5, 6, 7, ... represent the indexes for mutants.

It is applicable to estimate μ_1, B_{gb} and $P(B)_{gp}$ with a sample-based approach, i.e. using all the samples in the batch or plate. However, such estimation can produce biased estimates for the screens with disruptive perturbations or sensitive phenotypes. Therefore, we use control-based normalization, and estimate $\hat{\mu}_1$, $\hat{B}_{1b}$ and $\hat{P}(B)_{1p}$ by fitting the model in Eq. (2.13) to the first control (i.e., to biological samples with $g = 1$ in the notation above). On the other hand, adding all the negative control into the normalization model produces extra statistical factors due to the difference between control samples. Part of this extra variation goes into the model error and affects the accuracy of estimating the plate and batch effect. Consequently, we use one control sample instead of multiple control samples to avoid the extra variability that is not caused by the plate and batch effects.

In linear mixed models, such estimates are typically obtained by maximizing the restricted/residual maximum likelihood (REML) using Expectation-Maximum (EM) or Newton-

Raphson algorithms. The ridge-stabilized Newton-Raphson algorithm allows a faster convergence [35], and we use this algorithm as implemented in the R package `lme4`. The unbiased estimation of variances $\sigma^2_{P_1}$, $\sigma^2_{B_1}$ and $\sigma^2_{\varepsilon_1}$ are derived. We normalize the scored phenotype of mutants to the first control ($g = 1$) by subtracting its estimated batch- and plate-specific deviations $\hat{B}_{1b}$ and $\hat{P}(B)_{1p}$ from the scored phenotypes of mutants ($g = 5$, 6, 7, ...). In Eq. (2.14), the normalized phenotype for the k^{th} replicate of the g^{th} mutant located in the b^{th} batch and on the p^{th} plate is denoted as r_{gkbp}.

$$r_{gkbp} = X_{gkbp} - [\hat{B}_{1b} + \hat{P}(B)_{1p}], \tag{2.14}$$

Extensions. We develop the above-mentioned linear mixed-effect model as a flexible alternative that can be extended in various ways to account for within-plate effects, confounding effects, or time-dependent correlation effects. For example, if the position of a mutant at columns and rows within a plate systematically change the phenotype, then similar as B-score in Eq. (2.1), the row factor and column factor can be included in the normalization models shown as follows.

$$\begin{gathered} X_{ij,gkbp} = \mu_g + R_{ip} + C_{jp} + B_{gb} + P(B)_{gp} + \varepsilon_{gkbp}, \\ \textstyle\sum_i R_{ip} = 0,\ \sum_j C_{jp} = 0, \\ P(B)_{gp} \overset{iid}{\sim} \mathcal{N}(0, \sigma^2_{P_g}),\ B_{gb} \overset{iid}{\sim} \mathcal{N}(0, \sigma^2_{B_g}),\ \varepsilon_{gkbp} \overset{iid}{\sim} \mathcal{N}(0, \sigma^2_{\varepsilon_g}) \end{gathered} \tag{2.15}$$

where $X_{ij,gkbp}$ is the intensity of k^{th} replicate of g^{th} mutant on p^{th} plate nested in b^{th} batch, R_{ip} and C_{jp} are the estimated effects due to the i^{th} row and the j^{th} column on the p^th plate. The remaining notation is kept same as those in Eq. (2.13). When only a small number of distinct mutants [36] are included in rows or columns, a lowess-based smoothing of these effects can be used.

For another example, we can also extend the proposed model to account for confounding effects on the scored phenotypes, such as growth rate of the mutants. Firstly the nor-

malization model in Eq. (2.13)-Eq. (2.14) can be utilized to obtain the estimated effects due to batches and plates relative to the scored growth rate GR_{gkbp}.

$$\text{GR}_{1kbp} = \mu_1 + B_{1b} + P(B)_{1p} + \varepsilon_{1kbp} \tag{2.16}$$

$$gr_{gkbp} = \text{GR}_{gkbp} - [\hat{B}_{1b} + \hat{P}(B)_{1p}], \tag{2.17}$$

And then a linear model can be fit to estimate a single linear relationship between the normalized confounding factor gr_{gkbp} and the normalized phenotype of mutants across all the biological samples.

$$r_{gkbp} = \beta_0 + \beta_1 gr_{gkbp} + \epsilon_{gkbp},\ \varepsilon_{gkbp} \overset{iid}{\sim} \mathcal{N}(0, \sigma^2_{\varepsilon_g}) \tag{2.18}$$

The normalized phenotype that account for the confounding factors are obtained as

$$r'_{gkbp} = r_{gkbp} - \hat{\beta}_1 gr_{gkbp} \tag{2.19}$$

The normalized phenotypes are comparable across mutants. For the experimental datasets detailed in Chapter 3.4, because we did not identify substantial within-plate effects due to rows and columns in the step of quality control, the basic normalization in Eq. (2.13)-Eq. (2.14) and the adjustment for growth rate in Eq. (2.18)-Eq. (2.19) are used in this thesis.

2.3.3 Estimation of variation and summarization

Illustrated in Fig. 2.2 - Fig. 2.3 that are boxplots of the normalized Sulfur accumulation phenotype for three controls in hundreds of plates based on the knock-out screen (KO) described in Chapter 3.4, we argue that the extra variation remained in the normalized phenotype is required to be estimated and then applied into the summary score for each mutant.

We used the first control, shown in Fig. 2.2, to derive batch- and plate-specific changes of phenotype $\hat{B}_{1b}$ and $\hat{P}(B)_{1p}$. The left and right panels are the before normalization phenotypes X_{1kbp} and the after normalization phenotypes r_{1kbp}. In the right panel, the Plate-wise means are on a horizontal straight line. It means that the systematic between-batch and between-plate variation is removed from the first control.

The normalized phenotype for the other two controls in Fig. 2.3 are not used to estimate the normalization parameters. Although the normalization removed large artifacts, such as the outlying measurements in the left panel are adjusted and then a systematic horizontal pattern is obtained in the left panel, the normalization step do not entirely eliminate all between-batch and between-plate deviations. The differential effect of batches and plates on mutant phenotypes produces such residual variation. The source of the residual variation can be the non-additive effect of interaction factors (i.e. between batch and mutant, and between plate and mutant), as well as the uncertainty in estimation of $\hat{B}_{1b}$ and $\hat{P}(B)_{1p}$ upon a small number of replicates in a plate.

When all replicates of mutants are repeated in all plates, the interaction factors can be accounted for by including them as fixed effects into the normalization model in Eq. (2.13),

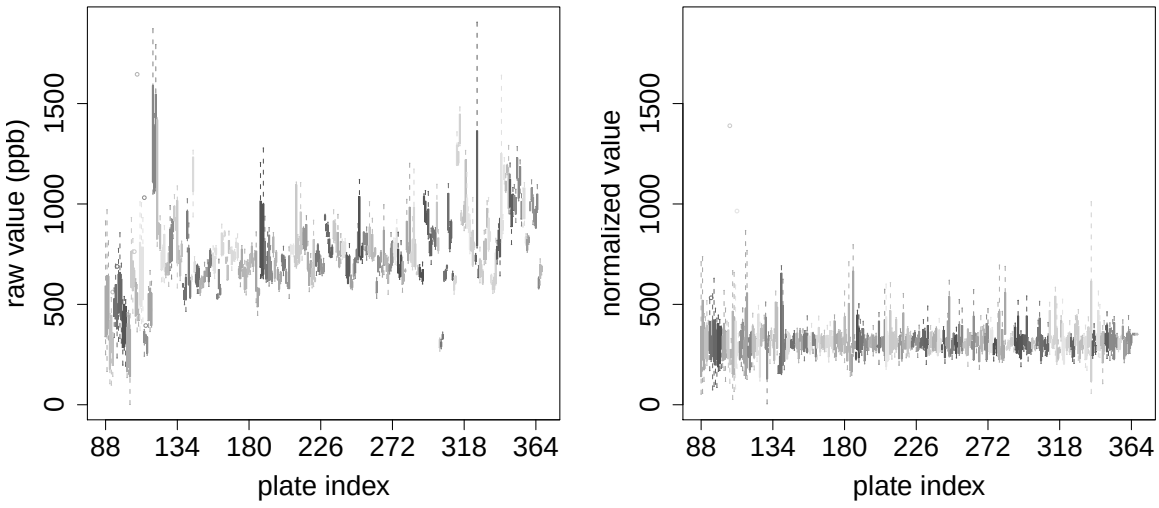


Fig. 2.2.: Effect of normalization on elemental abundance of Sulfur in the first control of the knock-out screen in Chapter 3.4, BY4741, which was used for normalization. The left is the boxplot of Sulfur abundance before normalization. The right is the boxplot after normalization. *Y axis:* raw or normalized abundance. *X axis:* plate id. The figures show boxplots of the phenotype in 305 plates, indicating batches by colors, similarly to panel (c) of Fig. 2.1.

or by using alternative approaches [22]. However, in large-scale perturbation screens where all replicates of the samples are profiled in a single plate instead of all plates, those inter-

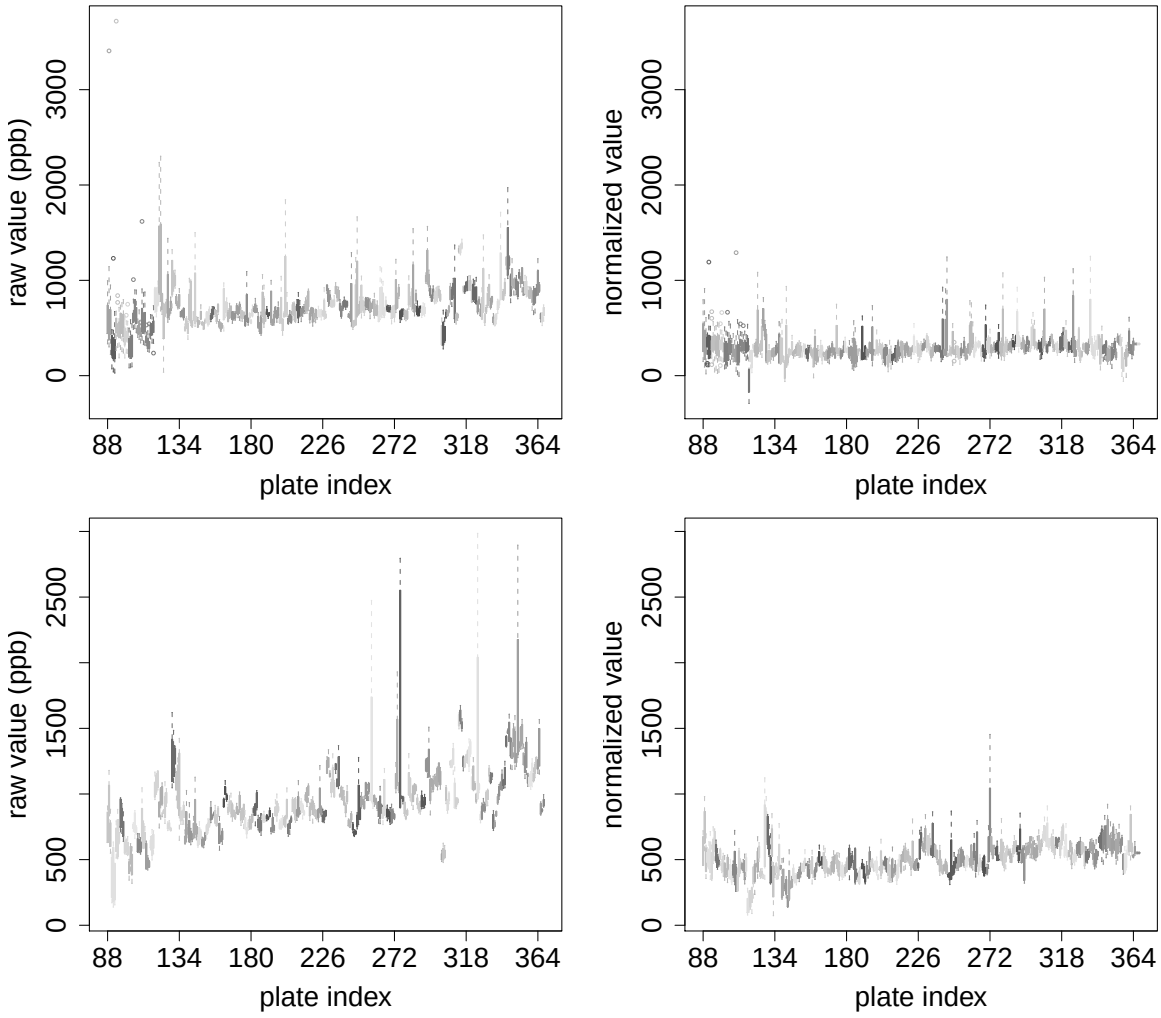


Fig. 2.3.: Effect of normalization on elemental abundance of Sulfur in the second (top row) and third control (bottom row) of the knock-out screen in Chapter 3.4, YDL227C and YLR396C, which are not used for normalization. The Y and X axes are the same as the axes in Fig. 2.2. The normalization procedure reduces the systematic differences between batches and plates, however still leave residual variation in the normalized values.

action effects cannot be estimated directly. Simply omitting the interaction effects can seriously underestimate the overall variation, and undermine the accuracy of the results.

Therefore, we propose to account for the residual variation in normalized phenotypes through the second linear mixed-effect model with random factors.

$$\begin{gathered} r'_{gkbp} = \mu'_g + P'(B)_{gp} + B'_{gb} + \varepsilon'_{gkbp}, \\ P'(B)_{gp} \overset{iid}{\sim} \mathcal{N}(0, \sigma^2_{P'_g}),\ B'_{gb} \overset{iid}{\sim} \mathcal{N}(0, \sigma^2_{B'_g}), \\ \varepsilon'_{gkbp} \overset{iid}{\sim} \mathcal{N}(0, \sigma^2_{\varepsilon'_g}),\ for\ g = 2, 3, 4, 5, ... \end{gathered} \tag{2.20}$$

where r'_{gkbp} is the normalized phenotype of sample g accounting for the confounding effect such as due to growth rate, and the remaining notation is same as in Eq. (2.13).

The summary phenotype of mutant g is represented by μ_g, and its estimate is equivalent to the average of the observed phenotypes over all replicates $\bar{r}'_g$ require all replicates for a mutant are nested in single plate. The associated estimated variation is

$$\widehat{Var}(\bar{r}'_g) = (\hat{\sigma}^2_{P'_g} + \hat{\sigma}^2_{B'_g} + \hat{\sigma}^2_{\varepsilon'_g}/n_g)$$

where n_g is the number of replicates for mutant g. The sample variance $s^2_{\varepsilon'_g}$ is used to estimate parameter $\hat{\sigma}^2_{\varepsilon'_g}$. However, since there are no across-plate replicates for mutants, $\sigma^2_{P'_g}$ and $\sigma^2_{B'_g}$ are not estimable directly.

Under the assumption that the control-based estimates accurately represent the residual variation of all the biological samples in the screen, we propose to utilize the information from one or several additional controls. In our practice the assumption can be plausible, and yields accurate results. In the situation where this assumption deviates from the characteristics of a screen, the experimental design has to be changed and the between-plate replicates need to be profiled. However the design substantially decreases the throughput and can be difficult to quantify large-scale perturbations in practice upon existing technologies.

These controls used for estimation of residual variance are not previously used in the normalization step. We obtain and plug-in estimates of $\sigma^2_{P'_g}$ and $\sigma^2_{B'_g}$.

$$\hat{\sigma}^2_{P'_g} + \hat{\sigma}^2_{B'_g} \approx \sigma^2_{P'_2} + \sigma^2_{B'_2} \text{ for } g = 3, 4, 5, 6, ... \tag{2.21}$$

When fitting Eq. (2.20) to the normalized phenotype of the second control such as shown in Fig. 2.3, the ratio between variance of factors and variance of model error is about $(\hat{\sigma}^2_{B'_2} + \hat{\sigma}^2_{P'_2})/\sigma^2_{\varepsilon'_2} \approx 0.23$ using the R package `HTSmix`. It measures the relative importance of the residual variation. Furthermore in Chapter 2.5.2 we illustrate that the changes in results using the proposed procedure is low when selecting different controls for normalization and variance estimation.

2.3.4 Determination of hits

Hypothesis and test statistic. The typical hypothesis for a mutant in perturbations screens $H_0 : \mu_g = \mu_0$ versus $H_0 : \mu_g \neq \mu_0$. The reference value of μ_0 can be obtained based on controls to test H_0 : Phenotype of mutant g is consistent with the phenotype of controls against H_a : Phenotype of mutant g is systematically larger (or smaller) than the phenotype of controls [10, 15].

However, when the majority consists of disruptive perturbations or sensitive phenotypes in the experiments, an unpractically large number of hits can be produced in testing. In this case, we consider the reference value μ_0 with sample-based approach. The test hypotheses are verbalized as follows.

$H_0 :$ μ_g is consistent with the median phenotype of all mutants
$H_a :$ μ_g is systematically larger or smaller than the median phenotype of all mutants

We calculate the test statistic shown in Eq. (2.22) that measures the normalized phenotype in the units of its estimated standard deviation.

$$D_g = \bar{r}'_{g\cdot\cdot} / \sqrt{(\hat{\sigma}^2_{P'_2} + \hat{\sigma}^2_{B'_2} + s^2_{\varepsilon'_g}/n_g)} \tag{2.22}$$

Compared to the regular (or moderated) T-statistic, the denominator in Eq. (2.22) $\hat{\sigma}^2_{P'_2}$ and $\hat{\sigma}^2_{B'_2}$ incorporated by D_g, is more conservative. Therefore we are able to obtain fewer hits. The sampling distribution of D_g is approximately Normal when assumptions in Eq. (2.13), Eq. (2.20) and Eq. (2.21) are satisfied. The average of D_g over all mutants is not necessarily

0 but very close to zero. The nature of the effects in the screen decides the center and the scale of the distribution.

Controlling FDR in the list of hits. We apply Efron's approach in [30] to produce a list of hits while controlling the False Discovery Rate (FDR). The assumption of Efron's approach is that the test statistics D_g with non-systematic changes are generated from the same distribution. A mixture of distributions under H_0 and H_a can be modeled by the sampling distribution of D_g. According to [30], we further standardize D_g to ensure that its sampling distribution under H_0 can be represented by the Standard Normal distribution with mean as 0 and standard deviation as 1, i.e.

$$Z_g = \frac{D_g - median(D_g)}{median(\ |D_g - median(D_g)|\ ) \cdot C} \tag{2.23}$$

where the normalizing constant required for a robust unbiased estimation of the scale [37] is $C = 1/\Phi^{-1}(3/4) \approx 1.4826$. The approach in [30] is implemented with the R package `locfdr`. A Normal distribution is fitted to the center of the histogram of Z_g, and the two-side cutoff of Z_g can be obtained to control the FDR under 0.05.

The distributions of Z_g for multivariate phenotypes are comparable across dimensions. Therefore, we can combine all dimensions in a single distribution to optimize the quality of fit. When the assumptions in Eq. (2.13), Eq. (2.20) and Eq. (2.21) are satisfied, the distribution of Z_g under H_0 is approximately Standard Normal. In Fig. 2.4, it is illustrated by the three different perturbation experiments that the sampling distributions of Z_g are close to normal and no gross departures from the assumptions are observed.

2.4 Experimental datasets

Perturbation screens. The performance of the proposed approach is illustrated using three large-scale genetic perturbation screens of *S. cerevisiae* (baker's yeast). 1) The fully knock-out (KO) data involves the 4940 open reading frames in haploid cells that were fully silent one-by-one. 2) The partially knock-out (KOd) data involves 1127 open reading frames in diploid cells, in which one of the two copies of the gene was silent so that the

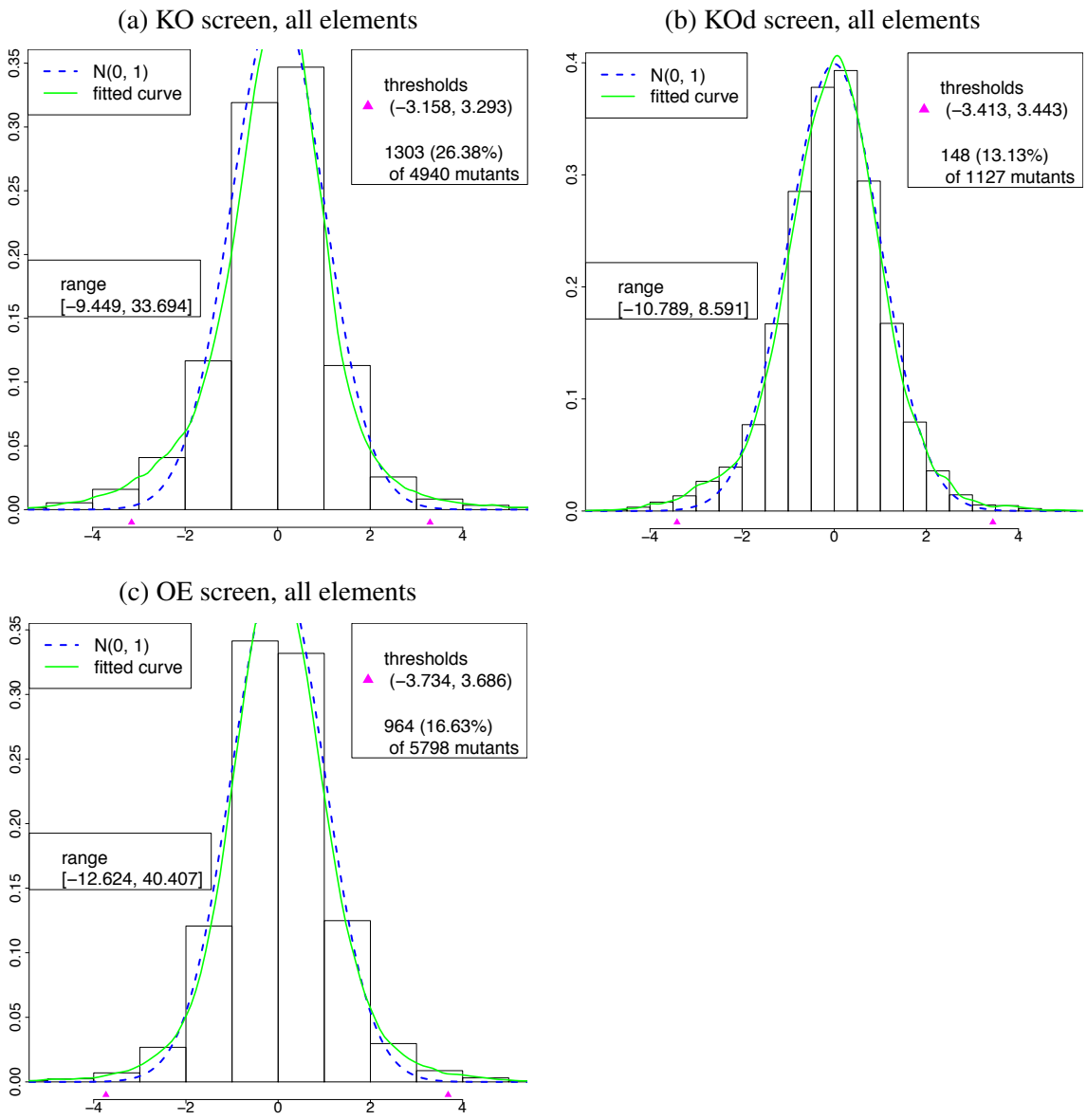


Fig. 2.4.: Determination of hits in three perturbation screens in the main manuscript. The histograms show the sampling distributions of Z_g in Eq. (2.23), combined across all dimensions of the multivariate phenotype. The dashed line showed the Standard Normal distribution fitted to the center of the distribution, and the green line shows the fit to the histogram based according to the two-group model in (Efron, 2008). Magenta triangles indicate the thresholds of Z_g, which control the FDR at 0.05.

perturbed cells still survive and grow. These mutant lines correspond to lethal disruptions in the haploid lines. 3) The over-expressed data OE involves the full collection of 5770 viable mutants in which each of the open reading frames is expressed at a higher than normal level. For all the three experiments, the mutants were incubated in a series of 96-well plates, with 4 (rarely 8 or 16) replicates per strain.

Abundance of the yeast ionome is the phenotype of interest in these screens. Mineral nutrient and trace element composition are the ionome of an organism [38, 39], including macronutrients (i.e. P, Ca, K, Mg); micronutrients essential for plant and human health (i.e. Cu, Fe, Zn, Mn, Co, Ni, Se, Mo, Cl); and minerals causing agricultural, environmental or health problems (i.e. Na, As, and Cd). To quantify each element [2], “a common yeast growth media was supplemented with additional elements, and each sample was processed, in batches of 3 plates, using inductively coupled plasma spectroscopy combined with mass spectroscopy (ICP-MS). Peaks in the spectra were signal-processed, and the absolute quantification in parts per billion (ppb) obtained through the use of calibration standards as described in [40]. A quality control procedure removed failed and outlying samples. Overall, the KO and KOd screen yields the multivariate phenotype of 14 elements, and the OE screen yields the multivariate phenotype of 17 elements for each mutant."

Two negative and two positive control strains are included by each of the three experiments, which are 1) BY4741, YDL227C, YLR396C and YPR065W for the KO screen, 2) BY4743, YDL227C, YLR396C and YPR065W for the KOd screen, and 3) YMR243C, YDL227C, YBR290W and YGL008C for the OE screen. Four replicates for each control strains were repeatedly add on all the plates. We selected the positive controls according to the results in [41]. The observable changes were found in key elements such as Ni60, Cd111 and S34 for these selected positive strains. They are helpful for us to test the ability of identifying such known changes in phenotype.

In the step of quality control for the ionomic profiles , we did not identify any significant spatial effect due to rows and columns within-plate effects detailed in Chapter 2.5.1. However, "it was established that differences of growth rates between mutants could act as potential confounders of the ionomic phenotypes" [2]. To account for the effect due to

different growth rates, all the samples and controls were quantified by the sample optical density (OD) with an OpsysMR plate reader (DYNEX Technologies, Chantilly, VA, USA). They are publicly available at www.ionomicshub.org.

A large number of mutations is expected to affect the ionomic phenotype since the elements constitute an integral part of most biochemical processes. Identifying the genes that has significantly stronger impact on at least one element abundance than its median overall all mutants help us discover the most important functional compounds.

2.5 Results

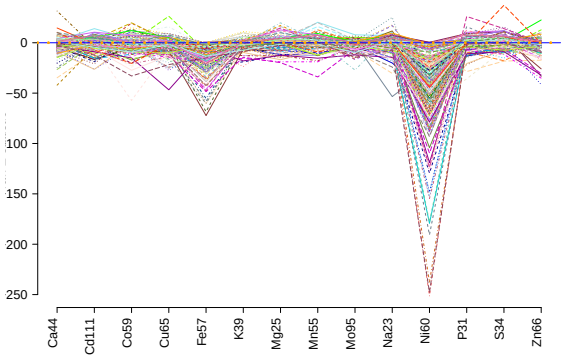
2.5.1 Rows and columns of the plate have negligible effect on the quantitative ionomic phenotypes.

We fit the additive model Eq. (2.1) to all the samples in a plate, separately for each plate. The quality control metrics for the plate are then defined as the median absolute deviation (MAD) of model-based estimates $\hat{R}_{i,p}$ and $\hat{C}_{j,p}$ relative to the median absolute deviation of the residuals r_{gkp} in Eq. (2.2).

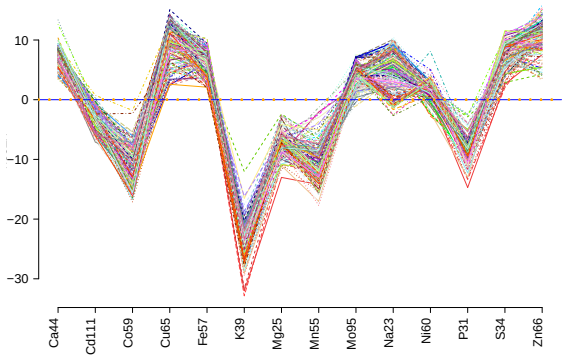
$$\text{MAD}_{ip}(\hat{R}_{i,p})/\text{MAD}_{ijp}(r_{gkp}) \quad \text{and} \tag{2.24}$$

$$\text{MAD}_{ip}(\hat{C}_{j,p})/\text{MAD}_{ijp}(r_{gkp}) \tag{2.25}$$

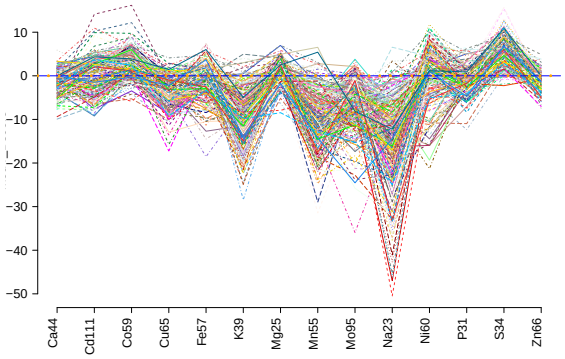
The boxplots shown in Supplementary Materials in [2] visualized the distributions of the quantities in Eq. (2.24) and Eq. (2.25) over all plates, separately for each inorganic element and each ionomic screen. Similar to row and column effects, and medians of the boxes vary deviations within 1. This indicates that the row and column effects may include the interesting phenotype variation, and the artificial effect due to rows and columns are not significant. Therefore, the column and row factors for Eq. (2.15) were not included for the three ionomic screens.



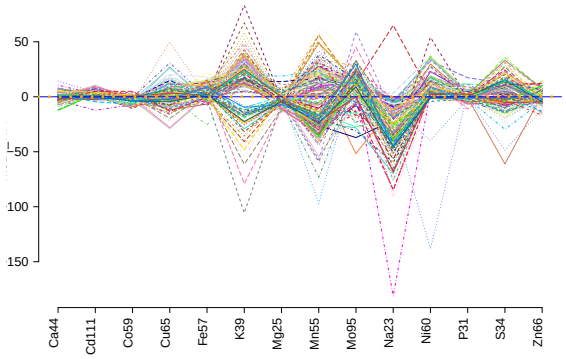
(a) Before normalization. Average: 0.558.



(b) After normalization using proposed methods. Average: 0.968.



(c) After normalization with B score and standardization with moderated T. Average: 0.640.



(d) After normalization with NPI and standardization with moderated T. Average: 0.438.

Fig. 2.5.: Profile plots of the standardized phenotypes of the control YLR396C in the KO screen, which has not been involved into normalization or standardization. *X axis:* inorganic elements. *Y axis:* (a) raw and (b)-(d) normalized and standardization phenotypes. Each line represents the phenotype of the control in one plate. In each profile plots, there are 305 lines corresponding to the phenotype of a control independently quantified in 305 plates.

2.5.2 Evaluation based on controls

The normalization procedure in Eq. (2.13) utilizes the information of the first negative control $g = 1$ (i.e. BY4741 for KO, BY4743 for KOd, and YMR243C for OE). The

procedure of estimating the variation in Eq. (2.20) is based on the second negative control $g = 2$ (i.e. YDL227C for KO, YDL227C for KOd, and YDL227C for OE). The other two

Table 2.1: Averages of Pearson correlations between profiles in pairs of plates after normalization and summarization. The values shown in the table were obtained based on the positive controls that were repeatedly measured in all the plates. They are not involved in normalization and estimation, and utilized for the purpose of evaluation. Higher values illustrates better noise reduction. (a) Normalization by B-score, standardization by Moderated T; (b) Normalization by Z-score, standardization by Moderated T; (c) Normalization by plate-wise median, standardization by Moderated T; (d) Normalization by Percent of positive controls, standardization by Moderated T; (e) Normalization by Percent of negative controls, standardization by Moderated T; (f) Normalization by Normalized percent inhibition, standardization by Moderated T; (g) Quantile normalization, standardization by Moderated T; (h) Proposed mixed-effect modeling for normalization with Eq. (2.13-2.14, 2.18-2.19), and standardization with Eq. (2.20-2.23). The methods in (a) - (g) are detailed in Chapter 2.3.2

	Averages of pairwise Pearson correlations between plates					
	(1) KO screen		*(2)* KOd screen		*(3)* OE screen	
	YLR396C	YPR065W	YLR396C	YPR065W	YBR290W	YGL008C
Existing methods:						
(a) B score	0.640	0.720	0.889	0.825	0.491	0.331
(b) Z score	0.765	0.776	0.910	0.817	0.530	0.361
(c) Plate-wise median	0.738	0.819	0.915	0.835	0.595	0.481
(d) PocMean	0.666	0.670	0.875	0.729	0.626	0.523
(e) PocMed	0.765	0.806	0.896	0.834	0.554	0.424
(f) NPI	0.438	0.508	0.689	0.686	0.759	0.595
(g) Quantile	0.696	0.772	0.857	0.917	0.630	0.485
Proposed methods:						
(h) Mixed model	0.968	0.971	0.963	0.940	0.962	0.961

positive controls are used to evaluate the quality of the results, which are YLR396C and YPR065W for KO, YLR396C and YPR065W for KOd, and YBR290W and YGL008C for OE.

Normalization and variance estimation: univariate phenotypes. The boxplots (Supplementary 1 in [2]) similar to Fig. 2.2 − Fig. 2.3 were visually checked for controls in all screens, and for the phenotypes of elements, before and after normalization with Eq. (2.13-2.14, 2.18-2.19). Illustrated by those figures, we observed that the proposed methods are able to remove the systematic trend in element abundance and obtain horizontal pattern across plates.

Boxplots of the normalization results using B score (Supplementary 2 in [2]), Z score (Supplementary 3 in [2]), and Normalized Percent Inhibition (Supplementary 4 in [2]) were also presented. These methods also eliminated the systematic trends across plates. However, the scale of the phenotype are changed, and then it is not straightforward for a relative comparison within or between dimensions. The multivariate phenotypes are utilized to illustrate problem and shown in the following paragraph.

Normalization and variance estimation: multivariate phenotypes. The relative efficiency of noise reduction procedures can be additionally evaluated by multivariate phenotypes. Since the significant difference in normalized phenotypes between plates within the same control is not biological meaningful and indicates poor reproducibility, we utilize the pattern of standardized multivariate phenotypes within controls between plates to evaluate the results of normalization and variance estimation. As compared to the mean phenotype in each dimension, a tighter pattern indicates a better elimination of the residual batch- and plate-specific variation.

The profile plots of the standardized phenotypes in Fig. 2.5 are for one positive control in the KO screen. Fig. 2.5 summarize the between pattern of 14 dimensional phenotype for the control when gene YLR396C was knockout. Fig. 2.5 (a) is before normalization, (b) is after the proposed normalization with Eq. (2.13-2.14, 2.18-2.19), and standardization with Eq. (2.20-2.23), (c) is after sample-based normalization with B-score and standardization

with moderated T statistic, (d) is after normalization with control-based Normalized Percent of Inhibition (NPI) and standardization with Moderated T. It is observed that B-score and NPI, combined with the moderated T statistic and the reference value in null hypothesis is 0, produced noisy standardized profile. The between-plate variation exceeds the between-element variation. Most elements have the average abundance around zero. The averaged correlations of multivariate phenotype between pairs of plates are only 0.640 and 0.438, which are close to averaged correlations without using any normalization method, 0.558.

However, Fig. 2.5 (b) illustrates that the proposed normalization and estimation procedure produces the tightest pattern. The averaged correlation is as high as 0.968. The between-plate variation is much smaller than the between-element variation. This allows us to obtain repeatable result of detecting changes in element abundance. Supplementary Materials in [2] provide the other profiles plots for all the controls in the three screens using different methods.

To quantify the performance of a method, we calculate the averaged Pearson correlations of standardized profiles (as in Fig. 2.5) over plates within a control. Those values are listed in Table 2.3. Since the proposed approach produces the highest correlation, the noise is substantially reduced as compared to the other techniques.

Stability of noise reduction to choice of controls. In Table 2.2, the averages of pairwise Pearson correlations of profiles of the controls in the KO screen quantify the performance of proposed methods when varying the combination of controls. The averages are between 0.811 and 0.985 that yield consistently high correlations. Therefore, the proposed procedure produces the results with little sensitivity to the specific choice of controls.

Relative contribution of analysis steps to the overall accuracy is shown in Table 2.3. The average Pearson correlations of the validation controls changed dramatically when the proposed steps varied in the three screens. Normalization with respect to the covariate and estimation of residual variance terms ($\sigma^2_{B'}$ and $\sigma^2_{P'}$) contribute more to the noise reduction than the batch- and plate-wise normalization.

Table 2.2: Averages of Pearson correlations between plates when vary the combinations of controls for normalization and variance estimation, standardized as in Fig. 2.5(b). Rows: combinations of controls used for normalization and variance estimation. Columns: the controls that were not involved in the models

Normalization-Standardization	Evaluation samples, KO screen			
	BY4741	YDL227C	YLR396C	YPR065W
BY4741-YDL227C			0.968	0.971
BY4741-YLR396C		0.906		0.902
BY4741-YPR065W		0.985	0.968	
YDL227C-BY4741			0.966	0.960
YDL227C-YLR396C	0.811			0.838
YDL227C-YPR065W	0.979		0.970	
YLR396C-BY4741		0.980		0.974
YLR396C-YDL227C	0.974			0.973
YLR396C-YPR065W	0.975	0.979		
YPR065W-BY4741		0.977	0.966	
YPR065W-YDL227C	0.982		0.971	
YPR065W-YLR396C	0.857	0.881		

2.5.3 Evaluation based on mutant strains

Since under-estimating the between-plate variation is the main drawback of the existing procedures, the number of the resulting false positive hits can exceed the nominal FDR. To illustrate this problem, the moderated T statistics for the KO screen in Table 2.3 were considered. We fit the two-group model to determine the test statistic cutoff by controlling FDR not greater than 0.05. The number of mutants with at least one differentially abundant phenotype were determined. Results of such model fit using four existing procedures

Table 2.3: Pearson correlation of normalized and summarized profiles between plates, for two positive controls which have not been previously used for normalization or standardization. Higher values indicate better noise reduction. 'X' indicates the applied normalization and variance estimation steps. The first row corresponds to the proposed approach

			Average pairwise Pearson correlation between plates					
Normalize:		Estimate:	*(1)* KO screen		*(2)* KOd screen		*(3)* OE screen	
$B, P(B)$	growth	$\sigma^2_{B'}, \sigma^2_{P'}$	YLR396C	YPR065W	YLR396C	YPR065W	YBR290W	YGL008C
X	X	X	0.968	0.971	0.963	0.940	0.962	0.961
X		X	0.904	0.901	0.895	0.869	0.911	0.917
X	X		0.777	0.742	0.824	0.783	0.716	0.725
	X		0.831	0.797	0.743	0.739	0.705	0.684
X			0.202	0.207	0.176	0.243	0.160	0.292

are illustrated in Fig. 2.6. More results using the other existing procedures are shown in Supplementary Materials in [2].

The analysis resulted provided in [2] are "3497 (70%) hits using B score; 3709 (75%) hits using Z score; 4885 (98%) hits using Normalized Percent Inhibition; 4584 (92%) hits using Plate-wise median; 4044 (81%) hits using Percent of Positive Controls; 3962 (80%) hits using Percent of Negative control; 3359 (68%) hits using Quantile normalization. These numbers exceed the 1303 (26%) hits obtained using the proposed procedure, and likely contain some false positive hits." Some of these reduction in the number of hits may lead to a loss of sensitivity. However, the next section illustrates that the proposed approach is specific, and can help direct the follow-up experiments towards useful targets.

Detection of known changes in abundance. In [41], from the knockout library in yeast, 4358 mutants were assayed. The abundance of 13 elements were quantified. They are Ca, Co, Cu, Fe, K, Mg, Mn, Ni, P, Se, Na, S, and Zn. Inductively Coupled Plasma-Atomic Emission Spectroscopy (ICP-AES), was used for the quantification on the ionomic phenotypes. The instrument is less sensitive and subject to larger variation. Different controls

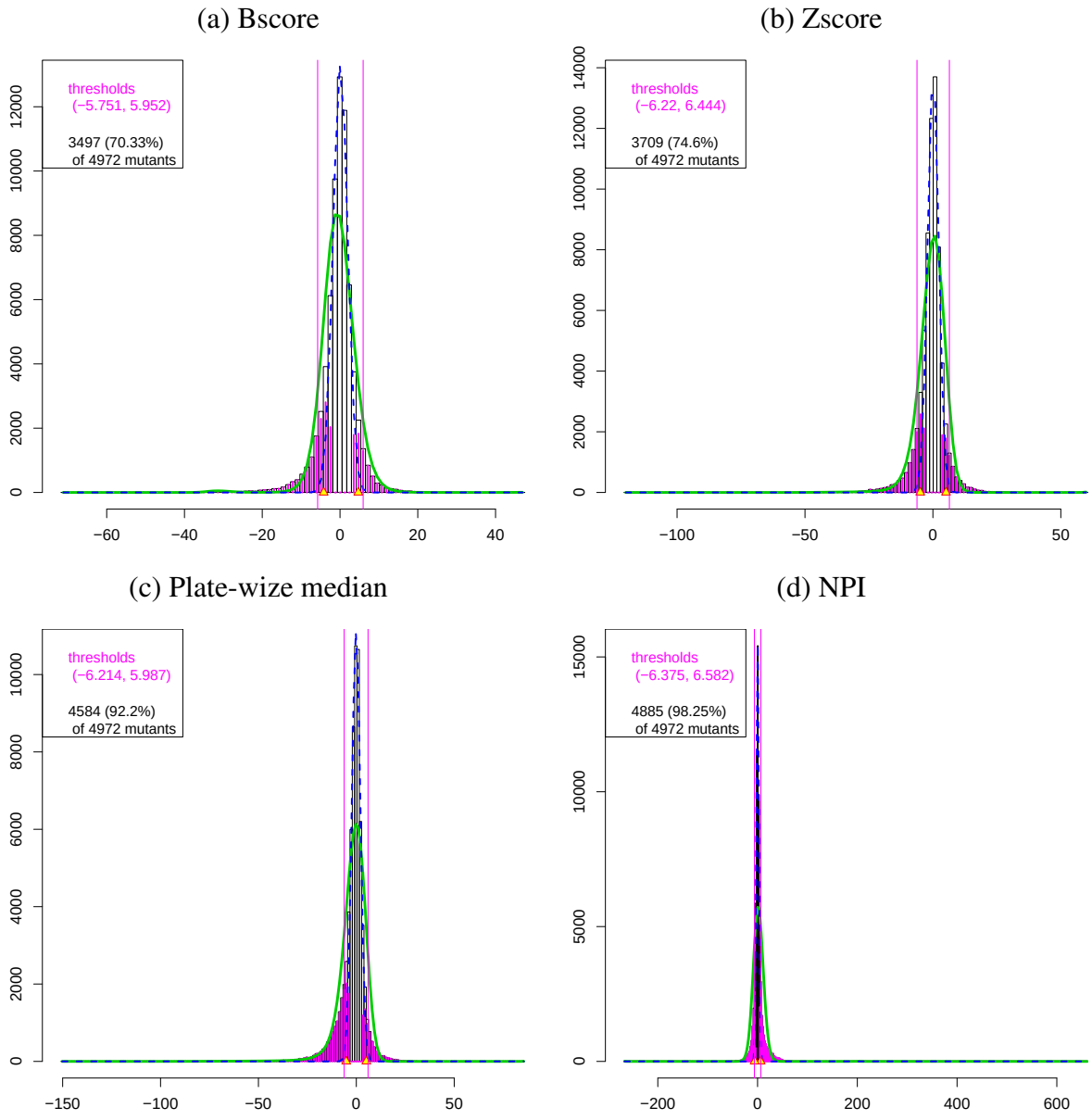


Fig. 2.6.: Result of fitting a two-group model by (Efron, 2008) to the test statistics of all mutants in the KO screen, combined across all the dimensions of the multivariate phenotypes. The raw phenotypes were normalized as described in legends (a-d), and standardized with the moderated T statistic. Score cutoffs were chosen to control the False Discovery Rate at 0.05.

were used, and no grow rate was adjusted. Given these differences in the experimental settings, 36 (i.e. 65%) of the KO hits reported by that study were confirmed in the list of hits identified by the proposed approach. Consequently, a sensitive detection of known changes in the phenotypes is enabled by the proposed noise reduction procedure.

Functional annotation of differentially abundant mutant strains. We further evaluate the specificity of the proposed approach by considering functional annotations of genes that statistically changed at least one element abundance. Annotations on those genes were obtained from the SGD database www.yeastgenome.org, and by literature search. As a result, 37 hits in the KO screen, and 19 hits in the OE screen, were described as involved in mineral regulation.

Three of those are YBR290W (BSD2Δ), YGL167C (PMR1Δ) and YPR194C (OPT2Δ). They were found as differentially changing the abundance of Cadmium (Cd) in the KO screen. The same conclusion on the involvement of these genes in Cadmium regulation has been previously established. In particular, endoplasmatic reticulum (ER)-localized membrane protein is encoded by BSD2 (bypass SOD deficiency). The protein controls the uptake of divalent metal ions from the growth medium [42]. Furthermore, the major Golgi membrane-localized Ca^{2+} and Mn^{2+}-transporting P-type ATPase is PMR1. It is essential for intracellular Cd^{2+} trafficking and detoxification [43, 44]. Finally, As an oligopeptide transporter, cells' sensitivities to anticancer drugs and divalent ion Cd can be increased by the loss-of-function of OPT2 [45].

Experimental validation. We also experimentally validated 19 KO mutant strains, which were determined as differentially abundant in Cd using the proposed design and analysis methods. *S. cerevisiae cells* in the validation experiment were grown overnight to an OD600 nm of 1.3. "Aliquots of the cell suspensions were then serially diluted 10-, 100- and 1000-fold and spotted onto solid YNB medium supplemented with the indicated concentrations of $CdCl_2$. Colonies were visually assessed after incubating plates for 2 days at 30°C." [2]

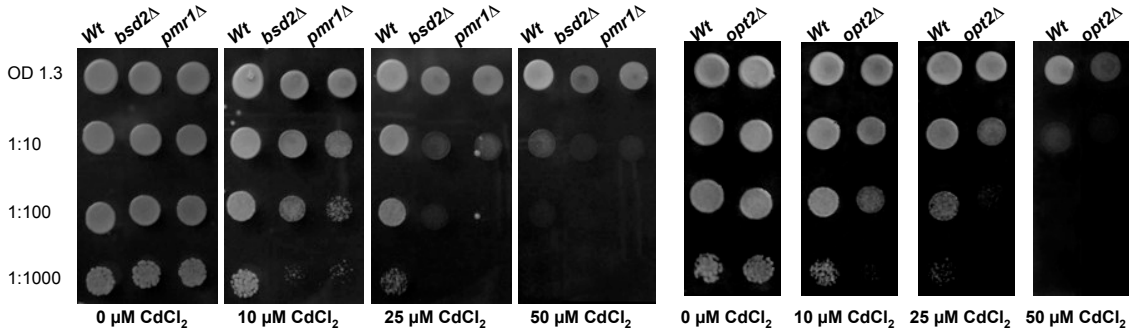


Fig. 2.7.: Cadmium sensitivity of BY4741 wild type (Wt) and selected mutant strains. (a) YBR290W ($BSD2\Delta$) and YGL167C ($PMR1\Delta$) to Cd supplement in growth medium. (b) YPR194C ($OPT2\Delta$).

The growth of three mutant strains among the 19 profiled mutants, YBR290W (BSD2Δ), YGL167C (PMR1Δ) and YPR194C (OPT2Δ) were compared to the wild-type BY4741 strain in the medium with or without Cd. Pictures of experimental results are shown in Fig. 2.7.

On the medium without Cd, the growth of the three KO strains is indistinguishable from the control strain. On medium supplemented with Cd, all the three KO strains are more sensitive to Cd than control strain. The sensitivity increased after Cd concentration was increased. In this validation experiment, we obtained the results that are consistent with the KO ionomic screen with the proposed methods. It was concluded that these lines accumulate more Cd than the majority of mutants. The direction of changing Cd abundance is also consistent with the existing literature. Evidence has been established for the role of BSD2, PMR1 and OPT2 in Cd detoxification [42, 43, 45]. We also obtained similar experimental confirmation for 18 out of the 19 mutant strains identified in the KO screen.

2.5.4 Extension results of biological conclusion

In a downstream research based on the results of the proposed methods, we further obtain meaningful biological conclusion in [3]. There are 1065 mutant strains finally identified with changes in element abundance. They are 584 mutant strains in KO screen, 35 mutant strains in KOd screens, and 446 mutant strains in OE screens. Using hierarchical clustering method, we identify important grouped roles for the mitochondria, vacuole and ESCRT pathway that regularize the ionome. Using network analysis, we detect hub genes such as PMR1 in Mn homeostasis. Combining the information curated by BioGrid [46], we identify novel members of ionomic networks such as SMF3 in vacuolar retrieval of Mn. All yeast ionomic data can be downloaded at www.ionomicshub.org.

2.6 Conclusion

Accurate interpretation of true biological phenotypes based on high-throughput perturbation screens is challenging due to restrictions on their experimental design and implementation. For example a plate can only have small number of samples, small number of replicates for each mutant have to be nested within plates, and then few or no mutant replication between plates. Consequently, effects due to plates, plates nested in batches, and confounding variable (i.e. growth rate) can dominate variation in scored phenotype.

In this work, we proposed an experimental design that requires at least two control samples. One is used for normalization and the rest are used for variance estimation based on linear mixed-effects models. We observed that control samples are less affected by outlying phenotype than mutant samples especially for the screens where a large proportion of samples show changes in the phenotypes. Motivated by this insight, we focused on control-based normalization, and substantially reduce those sources of artificial variation. We further account for the stochastic variation due to the interaction between plates and mutant samples. This step of estimating residual non-additive effects is important for the optimal detection of hits. By evaluating several existing procedures and the proposed procedures based on three perturbation screens profiling elements abundance, we illustrated in [2] that

“the proposed methods can obtain comparable scores used in conjunction with a practical experimental design, allows extensions to alternative structures of data, enables a specific discovery of biologically meaningful hits, strongly outperforms the existing approaches". Therefore this proposed procedures are recommended as a useful tool in high-throughput functional investigations for noise reduction and efficient detection of hits.

3. RNA-SEQ EXPERIMENTS

3.1 Introduction

[1] Whole transcriptome shotgun sequencing (RNA-seq) technology [47–49] quantifies gene expression in biological samples in counts of transcript reads mapped to the genes. Accurate and comprehensive, it has made a major impact on genomic research [50–52]. One common goal of RNA-seq experiments is to detect differentially expressed genes, i.e. genes for which the counts of reads change between conditions more systematically than as expected by random chance [51]. Statistical methods for detecting differentially expressed genes must reflect the experimental design, and appropriately account for the stochastic variation. Moreover, many RNA-seq experiments serve as high-throughput screens of a small number of samples with the goal of subsequent experimental validation. Therefore the analysis must handle a relatively small number of biological replicates.

A variety of statistical methods and software has recently been proposed for detecting differentially expressed genes. These include DESeq [53], edgeR [54, 55], baySeq [56], SAMseq [57], BBSeq [58]. We briefly overview these methods. We further propose a direct and effective approach for characterizing the variation in the counts of reads, which improves the sensitivity and specificity of detecting differentially expressed genes for experiments with small sample size. We support this approach with an open-source R-based software package sSeq.

[1] Copy rights are released with Publisher's permission.

3.2 Background

3.2.1 The Negative Binomial distribution

The input to the statistical analysis is a set of discrete counts of reads in each experimental run. Although the counts can be modeled with the one-parameter Poisson or Geometric distributions [59–61], it is often advantageous to use the two-parameter Negative Binomial distribution [53–56]. This distribution is more general and flexible, and can be viewed as a generalization of both Poisson and Geometric distributions.

A Negative Binomial random variable $X \sim \mathcal{NB}(p, r)$ counts the number of failures before the r^{th} success in a series of independent and identical Bernouilli trials with probability of success p. Its probability distribution is

$$P\{X = x \mid p, r\} = \begin{pmatrix} r+x-1 \\ x \end{pmatrix} (1-p)^x p^r \text{ for integer } x \geq 0 \text{ and } p \in (0,1), \quad (3.1)$$

$$E\{X\} = r\frac{1-p}{p}, \text{ and } Var\{X\} = r\frac{1-p}{p^2}$$

An alternative parametrization used in this manuscript is $X \sim \mathcal{NB}(\mu, \phi)$ such that

$$P\{X = x \mid \mu, \phi\} = \begin{pmatrix} \frac{1}{\phi}+x-1 \\ x \end{pmatrix} \left(\frac{\mu\phi}{\mu\phi+1}\right)^x \left(\frac{1}{\mu\phi+1}\right)^{\frac{1}{\phi}}, \quad (3.2)$$

$$E\{X\} = \mu, \text{ and } Var\{X\} = \mu + \mu^2\phi$$

In this parametrization

$$\mu = r\frac{1-p}{p} \text{ and } \phi = \frac{1}{r}; \text{ or equivalently } p = \frac{1}{\mu\phi+1}, \; r = \frac{1}{\phi} \quad (3.3)$$

Sum of Negative Binomial random variables is also Negative binomial random variable, that is, if $X_j \sim \mathcal{NB}(p, r_j),\ j = 1, \ldots, n$ and are independent, then $\sum_{j=1}^{n} X_j \sim \mathcal{NB}(p, \sum_{j=1}^{n} r_j)$.

There are several connections to the Negative Binomial distribution and the other distributions. Firstly, it generalizes the Poisson distribution for count data. If we assume that $X \sim Poisson(\lambda)$, and the expected value of the Poisson distribution is itself a random

variable $\lambda \sim Gamma(\phi^{-1},\ \mu\phi),\ \mu_{gi}, \phi_{gi} > 0$, then the marginal distribution of X (on average over all possible expected values λ) is

$$P\{X = x \mid \mu, \phi\} = \int_0^\infty P(X|\lambda) \cdot f(\lambda|\mu,\phi)\, d\lambda = \int_0^\infty \frac{\lambda^x e^{-\lambda}}{x!} \cdot \frac{\lambda^{\phi^{-1}-1} e^{-\lambda\mu^{-1}\phi^{-1}}}{(\mu\phi)^{\phi^{-1}}\,\Gamma(\phi^{-1})}\, d\lambda$$

$$= \frac{\Gamma(x+\frac{1}{\phi})}{x!\Gamma(\frac{1}{\phi})} \left(\frac{\mu\phi}{\mu\phi+1}\right)^x \left(\frac{1}{\mu\phi+1}\right)^{\frac{1}{\phi}} \tag{3.4}$$

This connection motivates the use of the Negative Binomial distribution in RNA-seq experiments. Marioni *et al.* [61] demonstrated that the Poisson distribution adequately represents the technical variation of the counts in replicated libraries of a same biological sample. The Negative Binomial distribution allows us to explicitly model the combined effect of the biological and technical variation, as it inflates the total variation beyond what is specified by the Poisson distribution.

The Poisson-Gamma model is not the only motivation for modeling counts of RNA-seq reads with the Negative Binomial distribution. The Negative Binomial distribution can also arise as the sum of Geometric distributions, i.e. if $X_j \overset{iid}{\sim} Geometric([1 + (\mu\phi)^{-1}]^{-1}),\ j = 1, \ldots, \phi^{-1}$, then $\sum_{j=1}^{\phi^{-1}} X_j \sim NB(\mu, \phi)$. Alternatively, the Negative Binomial distribution can be represented as a compound Poisson distribution, i.e. if $X_j \overset{iid}{\sim} Logarithmic\left(\frac{1}{1+(\mu\phi)^{-1}}\right),\ j = 1, \ldots, J$, and $J \sim Poisson\left(-\phi^{-1}\log\frac{1}{1+\mu\phi}\right)$, then $\sum_{j=1}^{J} X_j \sim NB(\mu, \phi)$.

We focus on the Negative Binomial distribution in what follows. Denote X_{gij} as the random variable that expresses the counts of reads mapped to gene g ($g = 1, \ldots, G$) in sample (or, equivalently, *library*) j ($j = 1, \ldots, n_i$) in condition i, and denote x_{gij} as the observed values. For simplicity we consider two conditions ($i = A, B$), however the discussion holds for pairwise comparisons of conditions in experiments with more complex designs. We are particularly interested in situations where n_A and n_B are small, e.g. 1-4. We consider the following parametrization:

$$\begin{aligned} &X_{gij} \sim \mathcal{NB}(\mu_{gi}, \phi_g), \text{ where } \mu_{gi} \geq 0,\ \phi_g \geq 0, \text{ such that} \\ &\mathrm{E}\{X_{gij}\} = \mu_{gi},\ \mathrm{Var}\{X_{gij}\} = \mu_{gi} + \mu_{gi}^2\phi_g\ (\overset{\text{denoted}}{=} V_{gi}) \end{aligned} \tag{3.5}$$

The *dispersion* parameter ϕ_g determines the extent to which the variance V_{gi} exceeds the expected value μ_{gi} [62, 63].

3.2.2 Motivation for the proposed approach

The estimation of Var$\{X_{gij}\}$ is the main focus of this work, and is based on the following considerations.

1. A naïve approach is to estimate Var$\{X_{gij}\}$ using the method of moments (i.e. the per-gene sample variance). However it is highly variable in experiments with a small sample size [64].

2. RNA-seq experiments simultaneously quantify the expression of many genes. The genes share aspects of biological and technical variation, and therefore a combination of the gene-specific estimates and of consensus estimates can yield better estimates of variation. Such approaches are increasingly popular with RNA-seq experiments [53, 55].

3. The variance of the Negative Binomial distribution is a known function of the expected value μ_{gi} and of the dispersion ϕ_g, where ϕ_g is gene-specific. Therefore an accurate estimation of the dispersion (e.g. by combining the gene-specific and consensus estimates, without explicitly modeling its relationship to μ_{gi}) can lead to an accurate estimation of the variance, while preserving the mean-variance relationship.

4. Finally, constraints of throughput, sample availability or cost may restrict the number of biological replicates. Although experiments with little or no biological replication have poor reproducibility and are undesirable, such under-replicated screens are the only practical option in some situations [65, 66]. They can only detect large changes in expression, and require an extensive downstream validation with complementary low-throughput experiments and large sample size. To detect differentially expressed genes in such screens we assume that the majority of the genes are not differentially expressed, and that for these genes the samples from all conditions can be viewed as

biological replicates [54, 67]. Under this assumption a consensus estimate of dispersion can help us to improve the accuracy of gene-specific estimates of variation.

Our main concern is in how to (1) accurately define the consensus estimate of dispersion, and (2) accurately combine the gene-specific estimates of dispersion with the consensus estimate.

3.2.3 Existing approaches for RNA-seq experiments

Among the existing methods, edgeR [54, 55], DESeq [53] and baySeq [56] assume the Negative Binomial distribution, and SAMseq [57] and BBSeq [58] utilize other flexible models. The approaches have been extensively evaluated [68], and are broadly used. Hardcastle and Kelly [56] found that the performance of DESeq, edgeR and baySeq is superior to that of DEGseq [69], Li and Tibshirani [60] found that SAMseq improves upon PoissonSeq [60]. We briefly overview these approaches in the historical order. Table 3.1 summarizes the discussion.

Probability model. *edgeR* models the count of reads with the Negative Binomial distribution. It includes normalization, which accounts for the changes in read counts due to technical artifacts such as different sequencing depth. The normalization factor can be the total library size (i.e. the number of reads in the library). A more accurate normalization factor is the 'effective' library size m_{ij}, which multiplies the size of the library ij by a robust estimate of the log-fold change of the total count in condition i as compared to a reference run [67]. The parameter p_{gi} in row (b) of Table 3.1 is the probability that a single read maps to gene g for a sample in condition i. The model assumes that the dispersion parameter is gene-specific, but constant across conditions. For experiments without replication versions up to 2.4.6 assumed a common dispersion in all the genes. The subsequent versions discourage unreplicated experiments. Finally, alternatives based on generalized linear models for the Negative Binomial response are available.

Table 3.1: Existing and proposed approaches for differential analysis of RNA-seq experiments with two conditions. (a) s_{ij} is the size factor for sample j in condition i as defined in [53]. μ_{gi} is the expected normalized expression of gene g for a sample in condition i. $\hat{\phi}_g^{MM}$ is the per-gene dispersion estimate using the method of moments in Eq. (3.12). (b) m_{ij} is the 'effective' library size. p_{gi} is the probability that a read in i maps to gene g. *Up to v2.4.6. (c) ϕ_{gi} is gene- and condition-specific dispersion. $\hat{\mu}_{gi}$ and $\hat{V}_{gi}$ can be estimated by the method of moments or by the Cox-Reid corrected Maximum Likelihood. (d) N_{ij} is the size of the library i from condition j. p_{gi} is as in (b). (e) p_{gi} is as in (b). N_{ij} is as in (d). β is the coefficient of the linear predictor associated with an indicator Z of conditions. Column 'Time' is the run time for the experimental datasets in Chapter 3.4 on a laptop computer

		Probability model	**Estimation of dispersion**	**Testing**	**n=1**	**Time**
Negative Binomial	(a) **sSeq (proposed)** *This manuscript*	$X_{gij} \sim \mathcal{NB}\left(s_{ij}\mu_{gi}, \phi_g/s_{ij}\right)$	$\hat{\phi}_g^{sSeq} = \delta\xi + (1-\delta)\hat{\phi}_g^{MM}$, where ξ is a common dispersion and δ is a weight	$H_0 : \mu_{gA}$=μ_{gB} Exact test	Yes	mins
	(b) **edgeR** *Robinson & Smyth, 2008*	$X_{gij} \sim \mathcal{NB}\left(m_{ij}p_{gi}, \phi_g\right)$	$\hat{\phi}_g^{edgeR}$ maximize linear combination of per-gene & common-dispersion conditional likelihoods	$H_0 : p_{gA} = p_{gB}$ Exact or GLM-based test	Yes*	mins
	(c) **DESeq** *Anders & Huber, 2010*	$X_{gij} \sim \mathcal{NB}\left(s_{ij}\mu_{gi}, \phi_{gi}\right)$	$\hat{\phi}_{gi}^{DESeq} = \left(\hat{V}_{gi} - \hat{\mu}_{gi}\frac{1}{n_i}\sum_j \frac{1}{s_{ij}}\right)/\hat{\mu}_{gi}^2$ $\hat{V}_{gi}$ is estimated as function of the mean	$H_0 : \mu_{gA} = \mu_{gB}$ Exact or GLM-based test	Yes	mins
	(d) **baySeq** *Hardcastle & Kelly, 2010*	$X_{gij} \sim \mathcal{NB}\left(N_{ij}p_{gi}, \phi_g\right)$ Empirical priors on sets of parameters	$\hat{\phi}_g^{baySeq}$ maximize per-gene integrated quasi-likelihood	$H_0 : p_{gA} = p_{gB}$ Posterior probability cutoff	Yes	hours
Other	(e) **BBSeq** *Zhou,Xia & Wright, 2011*	$X_{gij} \sim \mathcal{B}inom\left(p_{gi}, N_{ij}\right)$ $p_{gi} \sim \mathcal{B}eta$, $logitE\{p_{gi}\} = Z\beta$, V($p_{gi}$)=E($p_{gi}$) (1-E($p_{gi}$))$\phi_g$	$\hat{\phi}_g^{BBSeq}$ maximize per-gene marginal likelihood; is a free parameter or a function of the mean	$H_0 : \beta = 0$ Wald test	Yes	hours
	(f) **SAMseq** *Li & Tibshirani, 2011*	Non-parametric	-	H_0: same distributions A&B Wilcoxon test & resampling	No	mins

DESeq also models the count of reads with the Negative Binomial distribution. It normalizes the read counts by a size factor s_{ij} [53]

$$\hat{s}_{ij} = \text{median}_g \frac{x_{gij}}{(\prod_{k=1}^{n_A} x_{gAk} \prod_{k=1}^{n_B} x_{gBk})^{1/(n_A+n_B)}} \tag{3.6}$$

The size factor can be thought of as the 'representative' ratio of counts in the library to the geometric mean of the counts in all the libraries, and differs from the 'effective' library size in edgeR. The parameter μ_{gi} in row (c) of Table 3.1 is the expected normalized expression of gene g in condition i. DESeq allows specification of separate variances for genes and conditions, and models the variances as functions of the expected values. This relationship can be a flexible smooth function (local polynomial) or a parabolic function $\hat{V}_{gi} = s_{ij} \cdot \hat{\mu}_{gi} + s_{ij}^2 \cdot \mu_{gi}^2 \cdot (a_0 + a_1/\hat{\mu}_{gi})$, where $a_0, a_1 > 0$ are constants. Alternatives based on generalized linear models for the Negative Binomial response are also available.

baySeq specifies the same probability model as edgeR, however it proposes a different Empirical Bayes characterization of the between-library variation. baySeq assumes that subsets of the libraries share the parameters of Negative Binomial distribution, and derives an empirical prior distribution for the corresponding parameter sets. After integrating over the empirical priors, the dispersion in the integrated likelihood is constant across conditions and different between the genes. The default normalization parameter is the library size.

BBSeq specifies a Beta-Binomial generalized linear model. Using the logit link, the model connects the expected probability of a read for gene g in condition i and sample j to the linear combinations of predictors, such as indicators of conditions and other covariates. The dispersion parameter can be independent from the mean (free model), or dependent on the mean (constrained model). Finally, *SAMseq* utilizes a fully non-parametric approach.

Estimation of dispersion. *edgeR* maximizes a weighted combination of the conditional log-likelihoods with per-gene dispersion and of the conditional log-likelihood with common dispersion. Conditional likelihoods generalize the restricted maximum likelihood (REML) estimation for a discrete response by conditioning on the sum of the read counts per class, and improve the statistical properties of dispersion estimates. The estimation requires calculating pseudocounts of reads that would have been obtained with libraries of

equal size, and an iterative computational optimization. For experiments with few replicates the estimates tend to be discrete values [70–72]. For experiments with many replicates, edgeR specifies a generalized linear model. Since conditional likelihoods cannot be easily extended to this case, these are further approximated by adjusted profile likelihoods [73].

DESeq starts by estimating per-gene means and variances of the normalized counts in each gene and condition by the methods of moments. Next, it re-estimates them by fitting the postulated relationship between the expected values and the variances. The estimates of dispersion can be back-calculated from the estimates of variance as shown in row (c) of Table 3.1. For experiments without replication, DESeq assumes that the majority of the genes are not differentially expressed, and combines the samples across conditions to estimate the variance. The same strategy is used with the generalized linear models.

baySeq relies on an iterative estimation of the relative gene expression and of the dispersion. Given an initial partition of the libraries into subsets and an initial estimate of the relative gene expression, it estimates the dispersion using the quasi-likelihood approach. Given the estimates of dispersion, it re-estimates the relative gene expression by maximizing the integrated likelihood. This is repeated for different partitions of the libraries into subsets.

BBSeq estimates the dispersion using maximum likelihood for the free model. For the constrained model it uses the estimates from the free model for all the genes, fits the postulated relationship to the mean, and re-estimates the dispersions. *SAMseq* side-steps the need to estimate the dispersion by using a fully non-parametric approach.

Testing. For the Negative Binomial model, *edgeR* tests the null hypothesis $H_0 : \ p_{gA} = p_{gB}$, and *DESeq* $H_0 : \ \mu_{gA} = \mu_{gB}$ separately for each gene. Both edgeR and DESeq utilize the exact test, which is free from asymptotic arguments and is therefore preferred. The test statistic for a gene is the total (normalized) count of reads in all the replicates of a condition. The p-value is the probability of the normalized read counts per group, such that under H_0 their probability is same or lower than the probability of the observed counts, conditional

on the total counts equal to the observed. With the generalized linear models, edgeR and DESeq use the asymptotic likelihood ratio or Wald tests.

baySeq ranks the genes according to their posterior probabilities of differential expression. *BBSeq* tests the coefficient of the linear predictor (i.e. condition) in the generalized linear model with the asymptotic Wald test. *SAMseq* utilizes a resampling strategy to estimate the distribution of the test statistic and the p-values.

3.3 Methods

The proposed approach combines aspects of the existing approaches, but is simpler, requires fewer assumptions and streamlines the computation. It is summarized in Table 3.1(a).

3.3.1 Probability model

Denote X_{gij} the counts of reads of gene $g = 1, \ldots, G$, replicate $j = 1, \ldots, n_i$ and condition $i = A, B$. Denote s_{ij} the size factor of the replicate j in the condition i. The probability model is

$$X_{gij} \sim \mathcal{NB}(\mu_{gi}\, s_{ij},\ \phi_g/s_{ij}),\ \text{where } \mu_{gi} \geq 0,\ \phi_g \geq 0,\ s_{ij} > 0 \text{ such that} \tag{3.7}$$

$$\mathrm{E}\{X_{gij}\} = \mu_{gi}\, s_{ij},\ \text{and } \mathrm{Var}\{X_{gij}\} = \left(\mu_{gi} + \mu_{gi}^2 \phi_g\right) \cdot\ s_{ij} \tag{3.8}$$

This model is one way to represent RNA-seq experiments, however it is quite flexible. First, the free per-gene dispersion parameter ϕ_g accommodates arbitrary dependencies of dispersion on the expected value, and is particularly useful in experiments with a small sample size where the true relationship may be obscured by the noise.

Second, the assumption regarding the size factors can be meaningful from the experimental viewpoint, and also for technical modeling reasons. From the experimental viewpoint, Eq. (3.8) shows that size factors s_{ij} linearly scale both the expected value of the counts of reads and their variance. Since the differences in library size are technical artifacts, and since the technical variation in RNA-seq experiments can be characterized with

the Poisson distribution, the linear scaling of the variance with the library size is consistent with the Poisson[2]. Moreover, this assumption enables us to use the Negative Binomial distribution to directly model both the counts of reads in one replicate library (necessary to introduce the size factors), and the sum of the counts in the replicate libraries of a condition (necessary to conduct the exact test). As detailed in Appendices, a traditional interpretation of a Negative Binomial random variable is 'the number of failures before the r^{th} success in Bernouilli trials with probability of success p'. In this traditional parametrization

$$X_{gij} \sim \mathcal{NB}(p_{gi}, r_g),\ j = 1, \ldots, n_i, \text{ such that } p_{gi} = \frac{1}{\mu_{gi}\phi_g + 1} \text{ and } r_g = \frac{1}{\phi_g}$$

Introducing the size factor as in Eq. (3.8) implies a constant p, i.e.

$$p_{gi} = \frac{1}{s_{ij}\mu_{gi} \cdot \frac{\phi_g}{s_{ij}} + 1} = \frac{1}{\mu_{gi}\phi_g + 1},\ r_{ij} = \frac{s_{ij}}{\phi_g} \tag{3.9}$$

Therefore, the size factors can be interpreted as scaling the required number of successes in the Bernouilli trials while fixing the probability of success. Then the sum of the counts

$$\sum_{j=1}^{n_i} X_{gij} \sim \mathcal{NB}\left(p_{gi}, \frac{\sum_{j=1}^{n_i} s_{ij}}{\phi_g}\right), \tag{3.10}$$
$$E\{\sum_{j=1}^{n_i} X_{gij}\} = \mu_{gi} \sum_{j=1}^{n_i} s_{ij}, Var\{\sum_{j=1}^{n_i} X_{gij}\} = \mu_{gi} \sum_{j=1}^{n_i} s_{ij} + \mu_{gi}^2 \sum_{j=1}^{n_i} s_{ij}\phi_g$$

We follow edgeR, DESeq and baySeq by specifying a Negative Binomial distribution. Since in experiments with a small sample size it may be difficult to distinguish the true dependency of dispersions on expected values from artifacts of random variation, the model specifies free gene-specific dispersion parameters ϕ_g. As the initial versions of edgeR we specify a common dispersion across conditions, i.e. ϕ_{gA}=$\phi_{gB} \overset{denoted}{=} \phi_g$. As a consequence, the counts of differentially expressed genes have different variances in each condition.

We follow DESeq in normalizing the counts by the size factor s_{ij}. However in the proposed normalization the size factor affects not only the expected value, but also the dispersion. Eq. (3.8) shows that under this assumption the size factor linearly scales both E$\{X_{gij}\}$ and Var$\{X_{gij}\}$. Such linear scaling is consistent with the technical variation in

[2]Alternative models of the scaling factors, such as $X_{gij} \sim \mathcal{NB}(\mu_{gi}\ s_{ij},\ \phi_g)$, assume extra-Possion scaling of the variance, i.e. Var$\{X_{gij}\} = \mu_{gi}\ s_{ij} + \mu_{gi}^2 \phi_g\ s_{ij}^2$. In experiments with small sample size it may be difficult to evaluate which model fits best, and the two approaches are similar when the size factors are close to 1.

RNA-seq experiments, which can be characterized by the Poisson distribution [61]. Since typical size factors are close to 1, the proposed model has little practical difference from the model in DESeq. However, as shown in Chapter 3.3.4, it allows us to directly conduct the exact test and contributes to the accuracy of the results.

3.3.2 Estimation of dispersion

Similarly to DESeq, we start by estimating the dispersion parameters by the methods of moments. A conservative estimate of the per-gene variance in experiments with a small sample size is obtained by pooling the samples across conditions, i.e.

$$\hat{V}_g = \frac{\sum_i \sum_j (x_{gij}/\hat{s}_{ij} - \hat{\mu}_g)^2}{\sum_i n_i \; - \; 1}, \text{ with } \hat{\mu}_g = \frac{\sum_i \sum_j x_{gij}/\hat{s}_{ij}}{\sum_i n_i} \tag{3.11}$$

and $g = 1, \ldots, G$. The estimate of dispersion $\hat{\phi}_g^{MM}$ is then calculated from Eq. (3.5), and negative values are truncated at zero

$$\hat{\phi}_g^{MM} = max\left(0, \; \frac{\sum_i n_i V_g - \mu_g \sum_i \sum_j \frac{1}{s_{ij}}}{\mu_g^2 \sum_i \sum_j \frac{1}{s_{ij}}}\right) \tag{3.12}$$

Unfortunately, in experiments with small sample size $\hat{\phi}_g^{MM}$ are unsatisfactory due to high variance [74–76]. Next we improve the statistical properties of these estimates by introducing shrinkage.

Stein [77] showed that when we estimate the expected values of three or more independent Normal random variables with known constant variance, shrinking the per-dimension estimates toward a target value ξ produces biased estimates, but reduces the mean squared error (MSE) for all choices of ξ. The shrinkage estimator by James and Stein [78–80] implements this strategy. More recently Hansen extended the approach of James and Stein with a generalized shrinkage estimator, [81]. Hansen's shrinkage can be used with any per-dimension estimator with an arbitrary sampling distribution (not necessarily Normal), for which the Central Limit Theorem holds. Specifically, it requires that the true parameter lies in a neighborhood of the restricted parameter space, and that the estimator is asymptotically Normal with a consistent variance. Estimators by the method of moments satisfy

these criteria. Applied to the estimation of ϕ_g, and assuming that the per-gene estimates are independent, the generalized shrinkage estimator is

$$\hat{\phi}_g^{sSeq} = (1-\delta)\hat{\phi}_g^{MM} + \delta \cdot \xi = \xi + (1-\delta)\left(\hat{\phi}_g^{MM} - \xi\right) \tag{3.13}$$

As can be seen, $\hat{\phi}_g^{sSeq}$ is a linear combination of the pre-defined target ξ and of the per-gene methods of moment estimates. The weight δ is defined as

$$\delta = \frac{\sum_g \left(\hat{\phi}_g^{MM} - \bar{\hat{\phi}}^{MM}\right)^2/(G-1)}{\sum_g (\hat{\phi}_g^{MM} - \xi)^2/(G-2)}, \text{ and } \bar{\hat{\phi}}^{MM} = \frac{1}{G}\sum_g \hat{\phi}_g^{MM} \tag{3.14}$$

Since $\sum_g \left(\hat{\phi}_g^{MM} - \bar{\hat{\phi}}^{MM}\right)^2 \leq \sum_g (\hat{\phi}_g^{MM} - \xi)^2$, the weight $\delta \in (0,1)$. Larger values of δ shrink the estimates closer to the pre-defined target ξ.

We utilize the Hansen's generalized shrinkage estimator $\hat{\phi}_g^{sSeq}$ in conjunction with the Negative Binomial distribution to test genes for differential expression. Although the assumption of $\hat{\phi}_g^{MM}$ being independent variables is simplistic, it is a suitable approximation for experiments with a small sample size. A similar assumption is made, e.g. by DESeq when modeling the variance as function of the mean. While the asymptotic argument cannot be justified in this context, we show empirically in Chapter 3.5 that $\hat{\phi}_g^{sSeq}$ performs quite well in practice.

Hansen showed that the estimator in Eq. (3.13)-Eq. (3.14) reduces the asymptotic MSE for all choices of targets ξ. However a good practice is to select a value for ξ that maximizes this reduction. To this end we approximate the MSE $= \mathrm{E}\{\sum_{g=1}^{G}(\hat{\phi}_g^{sSeq} - \phi_g)^2\}$ using the average squared difference (ASD) between $\hat{\phi}_g^{sSeq}$ and $\hat{\phi}_g^{MM}$

$$\text{ASD} = \tfrac{1}{G}\sum_{g=1}^{G}(\hat{\phi}_g^{sSeq} - \hat{\phi}_g^{MM})^2 \tag{3.15}$$

Eq. (3.15) substitutes ϕ_g with $\hat{\phi}_g^{MM}$, and divides MSE by the constant G for numeric stability. It is easy to show that ASD as function of ξ has the form

$$\text{ASD}(\xi) = \frac{\text{constant}}{\sum_{g=1}^{G}(\hat{\phi}_g^{MM} - \xi)^2} \tag{3.16}$$

Fig. 3.2(a) visualizes the functional form of $ASD(\xi)$ for a simulation in Chapter 3.4, and shows that the tail of the curve flattens for large ξ. Therefore we can also minimize the

bias by minimizing ξ, while enforcing the constraint that $ASD(\xi)$ is comparably small. In practice $sSeq$ estimates $\hat{\xi}$ by calculating the slope of $ASD(\xi)$, and setting

$$\hat{\xi} = \text{argmin}_{\xi} \{-\epsilon < \text{slope}(\text{ASD}(\xi)) < 0\} \tag{3.17}$$

for a small constant ϵ such as $\epsilon = 0.05$. The selected value is shown by the vertical line in Fig. 3.2(a).

Fig. 3.2(b) illustrates the fact that the proposed shrinkage estimator is a linear transformation of $\hat{\phi}_g^{MM}$. The slope of the transformation is $(1-\delta) \in (0,1)$, and the fixed point is the shrinkage target ξ. The shrinkage increases the per-gene estimates of dispersion that are smaller than ξ, and decreases the values that are larger than ξ. From our experience with multiple datasets, $\hat{\xi}$ is often around the 97.5^{th} quantile of $\hat{\phi}_g^{MM}$. In other words, it biases the majority of the estimates towards larger (and more conservative) values.

The proposed estimate of dispersion has analogies in methods developed for other high-throughput technologies. For example, it is similar in spirit to the moderated variance estimator in the package Limma [82, 83], which is also a linear combination of per-gene and consensus estimates.

3.3.3 Exact test for a two-group comparison

We follow edgeR and DESeq by testing $H_0 : \mu_{gA} = \mu_{gB}$ per gene with the exact test. The test statistic is $X_{gi\cdot}$, i.e. the sum of the read counts in each condition, $X_{gA\cdot} = \sum_{j=1}^{n_A} X_{gij}$ and $X_{gB\cdot} = \sum_{j'=1}^{n_B} X_{gij'}$. According to Eq. (3.10),

$$X_{gA\cdot} \sim \mathcal{NB}(s_A \cdot \mu_{gA},\ \phi_g/s_A), \text{ and } X_{gB\cdot} \sim \mathcal{NB}(s_B \cdot \mu_{gB},\ \phi_g/s_B) \tag{3.18}$$

where $s_A = \sum_{j=1}^{n_A} s_{Aj}$ and $s_B = \sum_{j=1}^{n_B} s_{Bj}$. Under H_0,

$$X_{gA\cdot} \overset{H_0}{\sim} \mathcal{NB}(s_A \cdot \mu_g,\ \phi_g/s_A), \text{ and } X_{gB\cdot} \overset{H_0}{\sim} \mathcal{NB}(s_B \cdot \mu_g,\ \phi_g/s_B) \tag{3.19}$$

Since the counts from the two conditions are independent, the joint probability distribution of $(X_{gA\cdot}, X_{gB\cdot})$ is

$$P\{X_{gA\cdot} = x_{gA\cdot}, X_{gB\cdot} = x_{gB\cdot}|H_0\} = P\{X_{gA\cdot} = x_{gA\cdot}|H_0\} \cdot P\{X_{gB\cdot} = x_{gB\cdot}|H_0\} \tag{3.20}$$

The p-value of the exact test is the combined probability of all the read counts per group, such that under H_0 they have a same or a lower probability than the counts observed, conditional on the total counts equal to the observed. Mathematically,

$$\text{p-value}_g = P_{2,g}/P_{1,g} \tag{3.21}$$

where

$$P_{1,g} = \sum_{x_{gA\cdot}, x_{gB\cdot} \in\ set1} P\{X_{gA\cdot} = x_{gA\cdot}|H_0\} \cdot P\{X_{gB\cdot} = x_{gB\cdot}|H_0\}$$
$$P_{2,g} = \sum_{x_{gA\cdot}, x_{gB\cdot} \in\ set2} P\{X_{gA\cdot} = x_{gA\cdot}|H_0\} \cdot P\{X_{gB\cdot} = x_{gB\cdot}|H_0\}$$

$$\begin{aligned} set1 &= \{X_{gA\cdot}, X_{gB\cdot} \mid X_{gA\cdot} + X_{gB\cdot} = x_{gA\cdot} + x_{gB\cdot}\} \\ set2 &= \{X_{gA\cdot}, X_{gB\cdot} \mid X_{gA\cdot} + X_{gB\cdot} = x_{gA\cdot} + x_{gB\cdot}\ \cap \qquad (3.22) \\ &\quad P\{X_{gA\cdot},\ X_{gB\cdot}\} \leq P\{x_{gA\cdot},\ x_{gB\cdot}\}\} \qquad (3.23) \end{aligned}$$

In practice the probabilities in Eq. (3.21) are calculated by substituting $\hat{\mu}_g$, $\hat{\phi}_g^{sSeq}$, $\hat{s}_A = \sum_{j=1}^{n_A} \hat{s}_{Aj}$ and $\hat{s}_B = \sum_{j=1}^{n_B} \hat{s}_{Bj}$ into the probability distributions in Eq. (3.19). The p-values are adjusted by method such as in [84] to control the False Discovery Rate.

3.3.4 Exact test for complex experiments

We consider two types of experimental designs that are more complex than a two-group comparison, and focus on pairwise comparisons of conditions.

Factorial experiments: The experiment in Hammer et al. [85] had a factorial design. It considered two factors (rat strains Sprague Dawley and L5 SNL Sprague Dawley 2, and time points 2 weeks and 2 months), and considered distinct biological replicates for each combination of strain and time.

Pairwise comparisons of conditions in such experiments is straightforward. We created a new condition with fours levels (Sprague Dawley, 2 weeks; Sprague Dawley, 2 months; L5 SNL Sprague Dawley 2, 2 weeks' L5 SNL Sprague Dawley 2, 2 months). The null hypotheses comparing pairs of conditions $H_0 : \mu_{gi} = \mu_{gi'},\ i, i' \in \{1, 2, 3, 4\},\ i \neq i'$, were

then tested as in Chapter 3.3.3, while the dispersion parameter were estimated using read counts from all the conditions.

Repeated measurements: The experiment in Tuch *et al.* [86] had a paired design, in that pairs of normal and tumor samples were obtained from each of the three patients. This is a special case of the repeated measurements design. The design is advantageous because it eliminates the between-subject biological variation from consideration when comparing the conditions.

To analyze such experiments we propose to view each subject as a separate independent unreplicated experiment. Since sSeq can handle unreplicated experiments, we derived the estimates of dispersion $\hat{\phi}_g^{sSeq}$ separately for each subject, thus reflecting the within-subject variation. Next we tested H_0 separately for each subject, obtaining separate p-values for each subject and each gene. Finally, assuming that the subjects are independent, we combined the p-values for each gene using Fisher's method [87] and obtained the consensus p-values.

There are at least two ways to obtain $\hat{\phi}_g^{sSeq}$. The first is to shrink the method of moment estimates $\hat{\phi}_{jg}^{MM}$ separately for each subject across conditions. The second is to average the per-subject method of moments estimates $\hat{\phi}_{jg}^{MM}$, and proceed with a single shrinkage step of the averaged estimates. We compared the sensitivity of these two approach for the Tuch dataset (detailed in Chapter 3.4). Fig. 3.1 shows that shrinking averaged estimates of dispersions resulted in a higher sensitivity. At the same time, both approaches are less sensitive than the analysis that ignores the paired nature of the design. The loss of sensitivity is possibly due to inefficiencies of the Fisher's methods, and several alternatives can potentially be considered [87].

Fig. 3.3 in the main manuscript is based on the shrinkage of the averaged method of moment estimates. The figure shows that despite the loss of efficiency the method has a same or a better accuracy of detecting differentially expressed genes as compared to the GLM-based approaches.

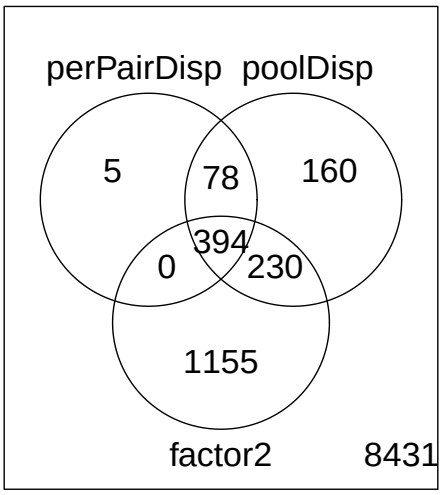


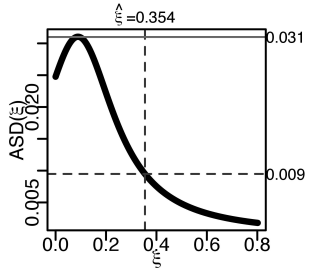
Fig. 3.1.: The number of differentially expressed genes in the Tuch dataset with paired experimental design. 'perPairDisp': separate dispersion estimation and shrinkage for each subject. 'poolDisp': averaged per-subject method of moments estimates of dispersion, and a single shrinkage step of the averaged estimates. 'factor2': analysis that ignores the paired nature of the design, and treats it as a two-group factorial experiment.

3.4 Datasets

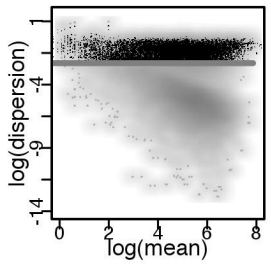
We evaluated the proposed approach using ten simulated and experimental datasets. The first five datasets had an external 'gold standard' of differential expression, and the last two had experimental designs more complex than a two-group comparison. **Simulation1, Simulation2** and **Simulation3** each generated G=20,000 genes in conditions A and B, $n_A = n_B = 2$. Parameters μ_{gA} and ϕ_g were simulation-specific (see below). 30% of the genes were simulated as differentially expressed, and for these genes $\mu_{gB} = \mu_g / exp(\epsilon)$ where $\epsilon \overset{iid}{\sim} \mathcal{N}(0.5,\ 0.25^2)$. Size factors were sampled from the Uniform distribution $s_{ij} \sim Uniform(0.5, 1.7)$, and rounded to the first decimal place. The simulated size factors are reported in Chapter 3.5.4. Read counts for gene g in sample j (j=1, 2) and condition i ($i = A, B$) were randomly generated from $\mathcal{NB}(\mu_{gi} \cdot s_{ij},\ \phi_g / s_{ij})$.

Fig. 3.2.: Dispersion and variance estimation in Simulation1. Similar plots for other datasets are shown in Supplementary Materials of [4]. (a) Average squared difference (ASD) versus shrinkage target ξ. ASD is maximized at $\xi = \bar{\hat{\phi}}^{MM}$ (solid horizontal line). The dashed lines are the selected target $\hat{\xi}$ and its ASD. (b) The proposed shrinkage estimator is a linear transformation of $\hat{\phi}_g^{MM}$, with the slope $(1-\delta) = 0.69$ and the fixed point $\hat{\xi} = 0.354$. All $\hat{\phi}_g^{MM} \leq 0$ are transformed to $\delta\xi = 0.11$. (c),(e) and (g) Dispersion estimates by sSeq, edgeR and DESeq, versus the per-gene mean read counts across conditions. Gray smooth scatter are $\hat{\phi}_g^{MM}$ (same on all the plots). Black dots are $\hat{\phi}_g$ estimated by each method. Gray lines indicate the true dispersion parameters. (d),(f) and (h) Same as above, but for the variances of the read counts

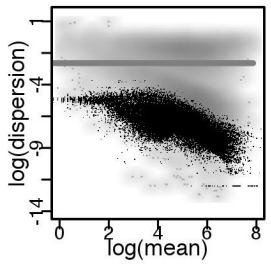
(a) Simulation1, n_i=2



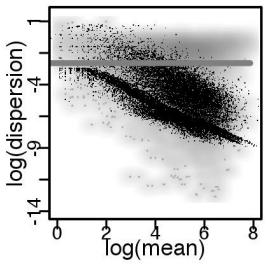
(c) Simulation1, n_i=2: Proposed



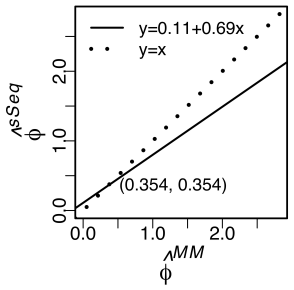
(e) Simulation1, n_i=2: edgeR



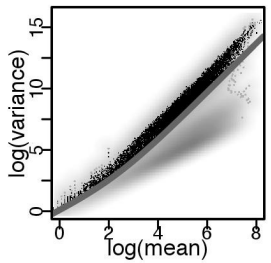
(g) Simulation1, n_i=2: DESeq



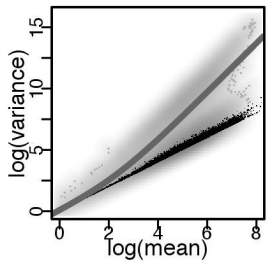
(b) Simulation1, n_i=2



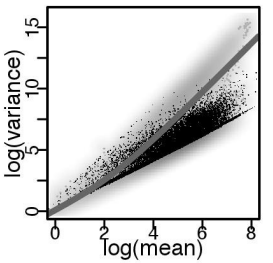
(d) Simulation1, n_i=2: Proposed



(f) Simulation1, n_i=2: edgeR



(h) Simulation1, n_i=2: DESeq



Simulation1 The expected values are randomly sampled from $\mu_g \sim$ Exponential($\lambda = 250$), where λ is the expected value. The dispersion parameter is considered as a constant across genes, $\phi_g = 0.1$.

Simulation2 As above, the expected values are randomly sampled from $\mu_g \sim$ Exponential ($\lambda = 250$), where λ is the expected value. The dispersion parameters are functions of the expected values $\phi_g = 1/(100 + \mu_g)$. This setting is the same as in [53].

Simulation3 From the dataset by Bottomly *et al* [88, 89], the largest experimental dataset in this manuscript, we selected a subset of non-differentially expressed genes (as determined by a consensus of sSeq, edgeR and DESeq), and sampled pairs $(\hat{\mu}_{gA}^{MM}, \hat{\phi}_g^{MM})$ from this subset as the true parameters.

MAQC [90] is the dataset from the MicroArray Quality Control (MAQC) consortium, comparing three libraries from Ambion human brain reference RNA against two libraries from Stratagene human universal reference RNA. The libraries were sequenced with the Illumina platform, resulting in 19,580 genes. The read counts for the RNA-Seq experiment were downloaded from http://www.ncbi.nlm.nih.gov, accession number SRA010153. The human genome `hg19` was downloaded from http://genome.ucsc.edu. The reads from each library were mapped to the human genome using Bowtie [91] command `bowtie -q -v 2 -a -m 1 -p 8 -quiet hg19 input.fastq output.map`.

A subset of the genes from four of the libraries were assayed by real-time reverse-transcription PCR (qRT-PCR) [90, 92]. To obtain the external 'gold standard' of differential expression, the qRT-PCR quantifications were downloaded from Gene Expression Omnibus (GEO, accession GSE5350). We compared the quantifications in the two conditions with the t-test. A gene was termed as differentially expressed if its p-value was less than 0.00001 (statistical significance), and the absolute fold change exceeded 2.1 (practical significance). A gene was termed non-differentially expressed if its p-value exceeded 0.2 (statistical significance), and absolute fold change was less than 1.5 (practical significance). This produced 323 differential genes and 85 non-differentially expressed genes. We used

the 323 differential genes and 85 non-differentially expressed genes determined by qRT-PCR as the 'gold standard'. Although the dataset only has technical replicates, it has been used extensively as the benchmark in the past [93–95].

Griffith *et al.* [96] compared fluorouracil (5-FU)-resistant human colorectal cancer cell lines MIP101 against their nonresistant counterpart MIP/5-FU24. The counts of aligned RNA-seq reads were downloaded from GEO (accession GSE23776). One library from each condition was quantified with the paired-end Illumina platform, resulting in 27,145 genes. 197 of these genes from the same samples were assayed by quantitative PCR (qPCR).

The qPCR data were downloaded from http://www.alexaplatform.org/alexa_seq/index.htm. As above, we compared the conditions with the t-test. For this dataset a gene was considered as truly differentially expressed if its p-value was less than 0.00001 (statistical significance) and the absolute fold change exceeded 3 (practical significance), and truly non-differentially expressed if its p-value exceeded 0.2 (statistical significance) and absolute fold change was less than 0.9 (practical significance). This produced 12 differentially expressed genes and 19 non-differentially expressed genes. We used 12 truly differential genes and 19 truly non-differentially expressed genes as determined by qPCR as the 'gold standard' for method comparison.

Brooks *et al.* [97] compared untreated cells of *Drosophila melanogaster* against cells cultured in presence of Pasilla, the homologue of the mammalian Nova-1 and Nova-2 protein. The table of read counts was downloaded from the R package **pasilla** published at Bioconductor. Two biological samples per condition were sequenced with the paired-end Illumina platform, resulting in 14,470 genes.

Sultan *et al.* [89, 98] compared two biological replicates of human cell lines Ramos B and HEK293T with the Illumina platform, yielding 6,573,643 uniquely aligned reads.

Bottomly *et al.* [88, 89] compared brain tissues of two inbred mouse strains, C57BL/6J (B6) and DBA/2J (D2), using the Illumina platform. The analysis of 10 and 11 biological samples per condition resulted in 343,445,340 uniquely aligned reads.

Hammer *et al.* [85, 89] compared gene expression in rat strains Sprague Dawley and L5 SNL Sprague Dawley 2, at two times (2 weeks and 2 months) in a factorial design. Two distinct biological libraries per condition and per time slot were quantified using the Illumina platform, resulting in 158,178,477 uniquely aligned reads.

Sultan, Bottomly and Hammer The read counts for these dataset were downloaded from http://bowtie-bio.sourceforge.net/recount [89].

Tuch *et al.* [86] compared the expression of genes in normal human tissues and in tissues with oral squamous cell carcinoma. The table of read counts was downloaded from GEO (accession GSE20116). The experiment had a paired design in that pairs of normal and tumor samples were obtained from three patient. The six libraries were sequenced using the SOLiD platform, resulting in 10,453 genes.

3.5 Results

We compared the performance of sSeq to the existing approaches. The following versions of the existing packages were used: edgeR v3.0.8 (January 2013), DESeq v1.10.1 (October 2012), baySeq v1.12.0 (October 2012), BBSeq v1.0 (March 2011), SAMSeq as part of the R package `samr` v2.0 (June 2011).

For sSeq, all the datasets were analyzed with the exact test, and analyses of the Hammer and the Tuch datasets accounted for their experimental designs. For edgeR and DESeq, the datasets with two-group comparisons were analyzed with the exact test, and Hammer and Tuch datasets were analyzed with the glm-based approaches. For edgeR, the `estimateCommonDisp` function in an older version of edgeR package (v2.4.6) was

Table 3.2: Areas under the ROC curves of detecting differentially expressed genes for the datasets with an external 'gold standard', while varying the FDR-adjusted p-value or posterior probability cutoff. Sub-columns are subsets of the data with one randomly selected replicate per condition, and the full datasets. Values closer to 1 indicate higher sensitivity and specificity

Methods		**Simulation1**		**Simulation2**		**Simulation3**		**MAQC Project**		**Griffith** *et al.*
		$n_i = 1$	$n_i = 2$	$n_i = 1$	$n_i = 2$	$n_i = 1$	$n_i = 2$	$n_i = 1$	n_A=3, n_B=2	$n_i = 1$
Proposed	sSeq	0.947	0.962	0.951	0.967	0.856	0.888	0.585	0.911	0.689
Existing	edgeR	0.918	0.948	0.938	0.951	0.840	0.833	0.558	0.850	0.557
	DESeq	0.932	0.940	0.937	0.949	0.842	0.816	0.577	0.867	0.596
	baySeq	0.568	0.711	0.548	0.714	0.558	0.628	0.551	0.852	0.702
	BBSeq	0.675	0.672	0.669	0.674	0.578	0.619	0.560	0.617	0.544
	SAMseq	-	0.964	-	0.968	-	0.882	-	0.563	-

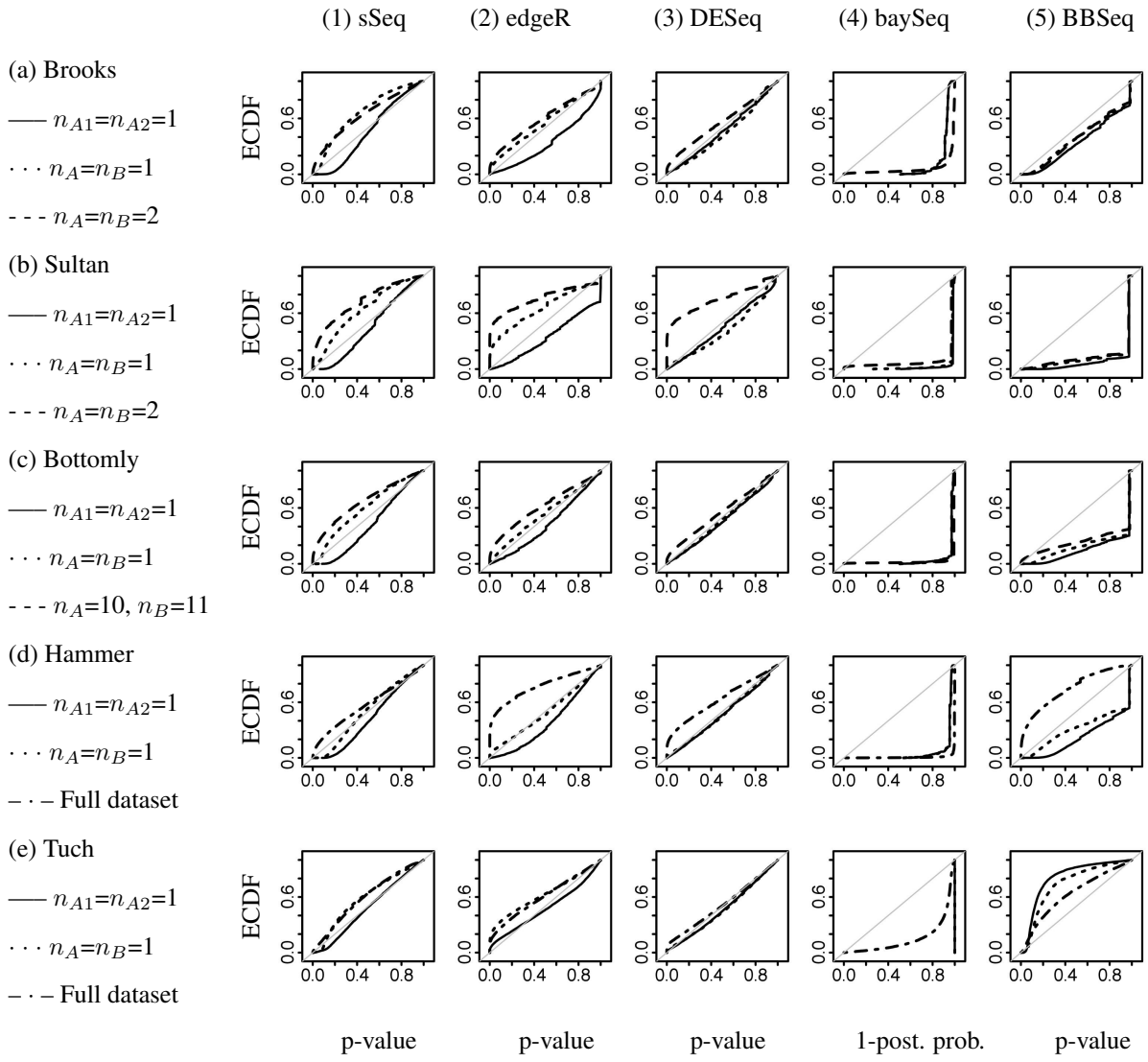


Fig. 3.3.: The empirical cumulative distribution function (ECDF) curves of detecting differential expression for the datasets with no external 'gold standard'. Y-axis: ECDF, function of the gene rank. X-axis: p-value or 1 minus posterior probability. Solid line: two randomly selected replicates from a same condition ($AvsA$). Dotted line: one randomly selected replicate from each condition (unreplicated $AvsB$). Dashed line: $AvsB$ on the full dataset for two-group designs. Dashed-dotted line: $AvsB$ on the full dataset for more complex designs. Gray line: 45 degree. SAMseq is not applicable to unreplicated experiments and is excluded. The desired patterns are high areas under the $AvsB$ curves, and $AvsA$ curves that are at or below the 45 degree line.

used to analyze un-replicated datasets. For DESeq, the option `fitType="local"` was used to estimate the per-group variance. The default parameters were used otherwise.

3.5.1 sSeq accurately estimates the variation

Since the proposed approach shares most similarities with edgeR and DESeq, we compared their estimates of dispersions and variances in more details. Fig. 3.2(c), (e) and (g) use the simplest case of Simulation1 to illustrate the estimates by the method of moments and by the three approaches. As expected, $\hat{\phi}_g^{MM}$ have a high variance, which increases with the mean. Also as expected, estimates by $\hat{\phi}_g^{sSeq}$ are biased towards larger values but have smaller deviations from the true values as compared to $\hat{\phi}_g^{MM}$. Estimates by the other two methods fit the pattern of $\hat{\phi}_g^{MM}$.

Fig. 3.2(d), (f) and (h) show that despite the differences in dispersion estimation, the estimates of variance by the three methods are less different. This is due to the fact that the values of the dispersions are small as compared to the means, and that the variances in Eq. (3.5) are highly influenced by the expected values. As the result, the bias in the estimation of the dispersion has a low impact on the overall estimation of variation. Similar plots for the other datasets are provided in Supplementary Materials of [4].

The first two columns of Table 3.2 show that the bias also has little impact on the performance of detecting differentially expressed genes, as the performance of sSeq, edgeR and DESeq are relatively similar. sSeq has a slightly higher area under the ROC curves.

Fig. 3.2(d), (f) and (h) also provide an insight into why shrinking the method of moments estimates of dispersion is more beneficial than shrinking the method of moments estimates of variance. The figures show that on the log scale the relationship between the mean and the variance in the Negative Binomial distribution is roughly linear for large mean counts. Mathematically, from Eq. (3.5)

$$\begin{aligned} \log(V_{gi}) &= \log(\mu_{gi} + \mu_{gi}^2\phi_g) = \log(\mu_{gi}) + \log(\mu_{gi}\phi_g + 1) \\ \log(V_{gi}) &\overset{\text{large } \mu_{gi}}{\approx} 2 \cdot \log(\mu_{gi}) + \log(\phi_g) \end{aligned} \tag{3.24}$$

A shrinkage of the variance estimates would multiply them by $(1-\delta) \leq 1$, and would distort the slope of the mean-variance relationship in Eq. (3.5) away from 2. The shrinkage of the dispersion parameter, on the other hand, preserves this nominal mean-variance relationship. Our results (shown in Supplementary Materials of [4]) confirmed that shrinking the variance leads to inferior performance.

To further investigate the usefulness of multiple shrinkage targets, we partitioned the genes into 10 groups according to the ranges of $\hat{\mu}_g^{MM}$, and applied the shrinkage separately to each group. Our results (not shown here) indicated that there is no advantage in specifying multiple shrinkage targets.

3.5.2 sSeq accurately detects differential expression

Five datasets with an external 'gold standard' were used to evaluate the sensitivity and the specificity of detecting differentially expressed genes. For each method the genes were ranked by FDR-adjusted p-value of posterior probability, and termed 'significant' for varying cutoffs. The sensitivity and the specificity of differential expression was compared to the 'gold standard', and summarized with Receiver Operating Characteristic (ROC) curves. Table 3.2 shows that the proposed approach consistently had a similar or a higher accuracy as compared to the existing methods.

Five datasets without an external 'gold standard' were used to evaluate the sensitivity and the specificity less formally, as discussed in [53]. First, comparisons of two conditions ('$AvsB$') had some truly differentially expressed genes. Therefore methods with higher sensitivity should have higher areas under the empirical cumulative distribution functions (ECDF) of the p-values defined as $\hat{F}(\text{p})=\frac{1}{G}\sum_{g=1}^{G} I_{\{\text{p-value}_g \leq p\}}$. Second, comparisons of replicates of a same condition ('$AvsA$') had no differentially expressed genes. Therefore methods with higher specificity should have ECDF curves at or below the 45 degree line. For baySeq we expect similar patterns of the ECDF curves based on the posterior probability cutoff. Fig. 3.3 summarizes the curves for the five datasets. It shows that sSeq produced

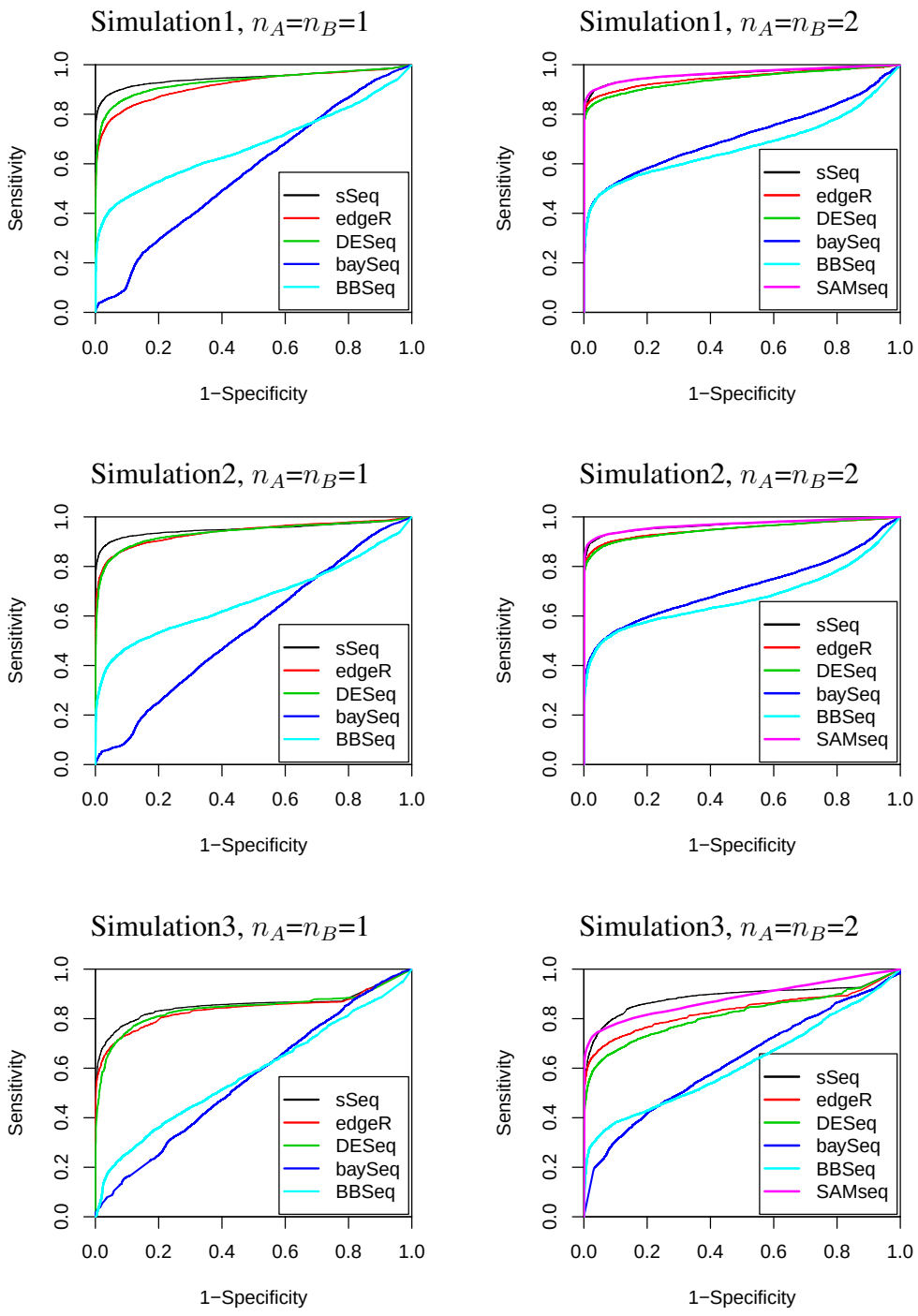


Fig. 3.4.: Areas under the ROC curves of detecting differentially expressed genes for the simulated datasets in Table 3.2.

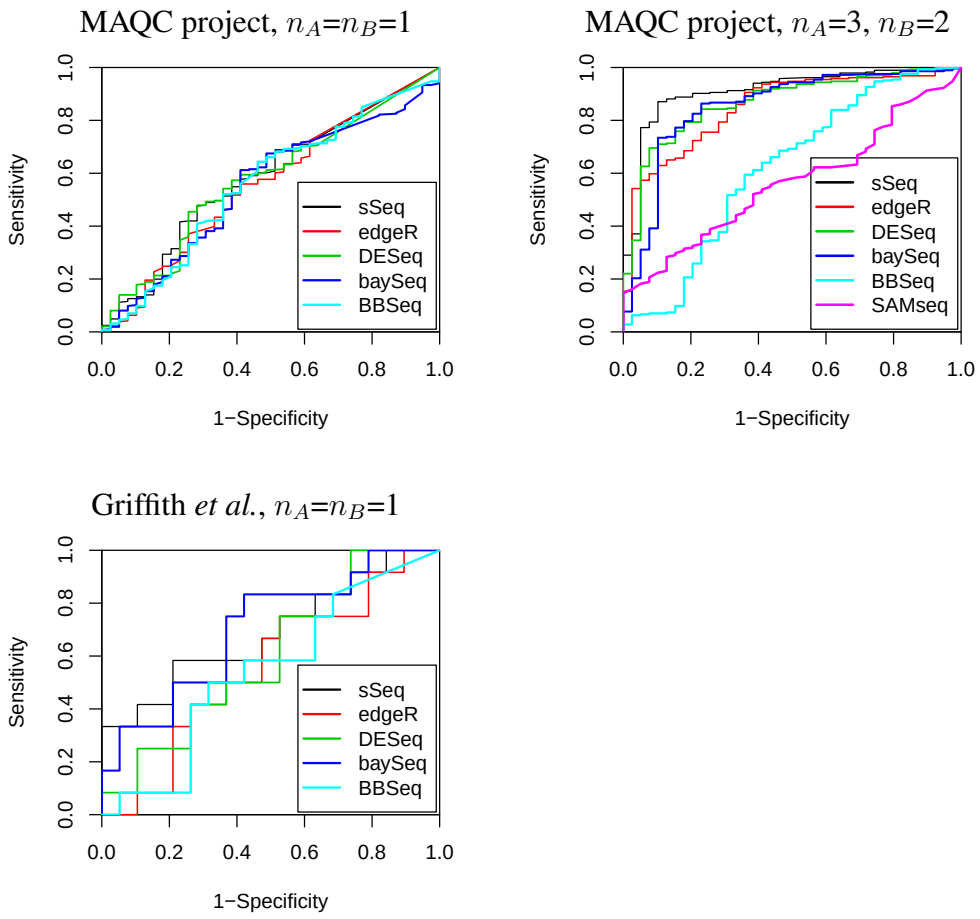


Fig. 3.5.: Areas under the ROC curves of detecting differentially expressed genes for the experimental datasets with an external 'gold standard' in Table 3.2.

Table 3.3: Areas under the ROC curves of detecting differentially expressed genes for the datasets with external 'gold standard', while varying the FDR-adjusted p-value or posterior probability cutoff, obtained with the **shrinkage of variance** as opposed to the proposed shrinkage of dispersion. Sub-columns are subsets of the data with one randomly selected replicate per condition, and the full available datasets. Values closer to 1 indicate higher sensitivity and specificity. The areas under the ROC curves are smaller than the values in the first row of Table 3.2

Simulation1		Simulation2		Simulation3		MAQC Project		Griffith *et al.*
$n = 1$	$n = 2$	$n = 1$	$n = 2$	$n = 1$	$n = 2$	$n = 1$	n_A=4, n_B=2	$n = 1$
0.605	0.863	0.654	0.885	0.602	0.506	0.597	0.522	0.646

most consistently the expected pattern, and had a similar or a higher accuracy as compared to the existing methods.

Table 3.3 and Fig. 3.6 show the details of detecting differentially expressed genes while shrinking the method of moments estimates of variance, as opposed to the proposed shrinkage of dispersion. They illustrate that shrinking the variance undermines the accuracy of the results.

3.5.3 Effect of sample size

To study the effect of sample size, we repeated simulation3 for $n_i = 1, 2, 3, 4, 6, 8, 10$. Table 3.4 summarizes the performance of the proposed method, as well as of DESeq and edgeR with both the exact test and the generalized linear model-based approach. The results indicate that sSeq is particularly advantageous for experiments with $n_i \leq 4$.

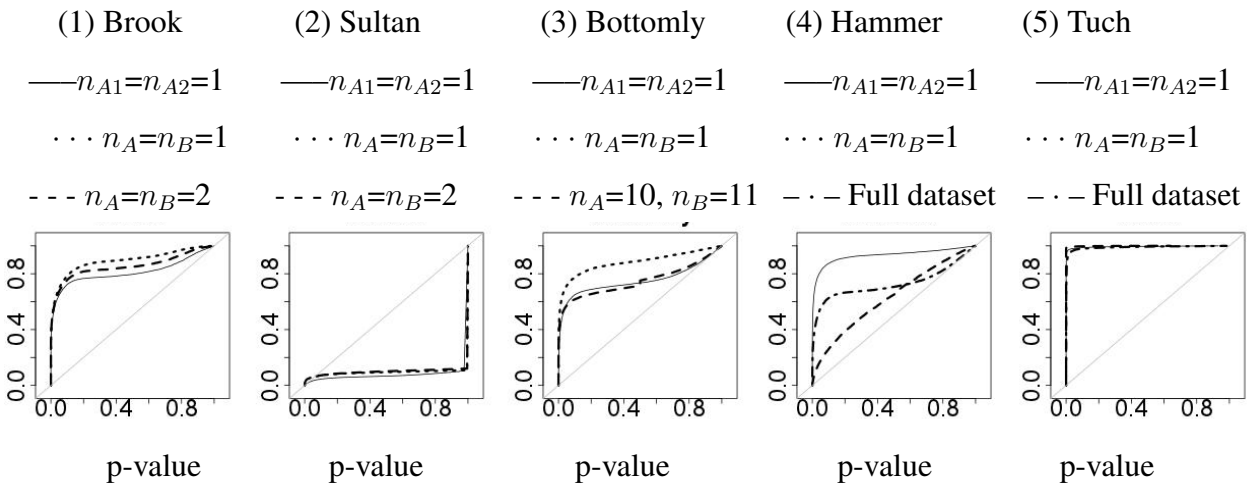


Fig. 3.6.: The empirical cumulative distribution function (ECDF) curves of detecting differentially expressed genes for the five datasets with no external 'gold standard' when shrinking the variance estimates. Y-axis: ECDF, function of the gene rank. X-axis: p-value. Solid line: unreplicated comparison $AvsA$. Dotted line: unreplicated comparison $AvsB$. Dashed line: $AvsB$ on the full dataset for two-group designs. Dotted-dashed line: $AvsB$ on the full dataset for more complex designs. Gray line: 45 degree. The curves are less consistent with the expected patterns than the curves in the first column of Fig. 3.2.

3.5.4 Effect of size factors

Table 3.5 the true and the estimated size factors for the ten datasets in this manuscript. The estimates are obtained with the proposed approach (i.e. are identical to the estimates by DESeq).

To investigate the effect of the estimates of size factors on the accuracy of the results, we conducted an additional evaluation for the three simulated datasets with $n_A = n_B = 2$, and for the methods that assume a Negative Binomial distribution. Table 3.6 shows the results of the original implementation of each method. Note that edgeR and baySeq estimate size factors on the total count scale, while the other methods use a relative scale. edgeR multiplies the total library size by the output of `calcNormFactors`. baySeq uses the total library size directly.

Table 3.4: Areas under the ROC curves of detecting differentially expressed genes for the Simulation3 when the number of samples increases. Values closer to 1 indicate higher sensitivity and specificity

$n_i =$	1	2	3	4	6	8	10
sSeq	0.856	0.888	0.903	0.913	0.917	0.926	0.929
edgeR	0.840	0.833	0.844	0.852	0.869	0.875	0.871
DESeq	0.842	0.815	0.885	0.894	0.907	0.914	0.915
baySeq	0.558	0.628	0.628	0.620	0.616	0.609	0.603
BBSeq	0.578	0.619	0.601	0.613	0.610	0.591	0.607
SAMseq	-	0.882	0.897	0.903	0.916	0.925	0.926
GLM edgeR	-	0.799	0.898	0.878	0.943	0.946	0.956
GLM DESeq	-	0.828	0.832	0.831	0.835	0.834	0.831

Table 3.7 uses the size factors estimated by sSeq (equivalently, by DESeq) with the two other methods. The size factors were converted to the appropriate scale by multiplying the total library sizes by the size factor estimates of sSeq and DESeq.

Table 3.8 uses the true size factors with all the methods. The size factors were converted to the appropriate scale for edgeR and baySeq by multiplying the total library sizes by the true simulated size factors.

To summarize, the results showed that size factors do indeed play a role in the accuracy of the results. Importantly, the changes did not affect the conclusions of the manuscript. The proposed method sSeq consistently showed a strong performance, except in Simulation3 where baySeq combined with a better size factor had a higher area under the ROC curve.

3.6 Discussion

In this manuscript we advocated a model that specifies free per-gene dispersion parameters in the Negative Binomial model for counts of RNA-seq reads. We also advocated

Table 3.5: The true and estimated size factors for the ten datasets. The estimates are obtained with the proposed approach (i.e. equivalent to DESeq). The true values of size factors are only available for simulated datasets

	s_{Aj}		s_{Bj}	
Datasets	**True**	**Estimated**	**True**	**Estimated**
Simulation1	0.6, 1.5	0.684, 1.729	0.8, 0.7	0.987, 0.866
Simulation2	0.6, 1.5	0.686, 1.732	0.8, 0.7	0.984, 0.863
Simulation3	1.3, 1.6	1.006, 1.196	1.4, 0.9	1.126, 0.749
MAQC		0.966, 1.202, 1.023		0.516, 1.706
Griffith *et al.*		1.317		0.760
Brooks		1.297, 1.042		0.819, 0.911
Sultan		0.917, 0.862		1.160, 1.132
Bottomly		between 0.578 and 1.524		between 0.756 and 1.616
Hammer		1.038, 0.897		1.027, 1.065
Tuch		0.719, 0.831, 1.753		0.627, 1.084, 1.424

a biased estimation of these parameters, which can reduce the variance of the estimates and minimize the overall mean squared error. Biased estimation is different from specifying a probability model (such as in DESeq) that assumes a true systematic relationship of the true variance and the true mean. It is particularly useful for experiments with a small sample size, where the systematic relationship may be difficult to evaluate. The shrinkage estimates are easy to compute, avoid iterative estimation, minimize the potential for overfitting, and do not require extra computation time. They are compatible with the exact test of differential expression. For the datasets in this manuscript, sSeq consistently had a similar or a higher sensitivity and specificity of detecting differential expression than the existing methods. The approach can be generalized to express the dependence of the dispersions on the expected value, or on other covariates such as GC content or Gene Ontology annotations.

Table 3.6: The estimates of the size factors by each method, and the corresponding areas under the ROC curves for the three simulated datasets

	Estimates $\hat{s}_{ij}$ by each method			**AUROC**		
	Sim1	**Sim2**	**Sim3**	**Sim1**	**Sim2**	**Sim3**
sSeq	0.684, 1.729, 0.987, 0.866	0.686, 1.732, 0.984, 0.863	1.006, 1.196, 1.126, 0.749	0.962	0.967	0.888
edgeR	2674477×1.119, 6677525×1.136, 4832029×0.887, 4230815×0.887	2651592×1.129, 6625427×1.130, 4818957×0.885, 4215156×0.886	7720510×1.132, 9168128×1.128, 10830341×0.885, 7196265×0.885	0.948	0.951	0.833
DESeq	0.684, 1.729, 0.987, 0.866	0.686, 1.732, 0.984, 0.863	1.006, 1.196, 1.126, 0.749	0.940	0.949	0.816
baySeq	2674477, 6677525, 4832029, 4230815	2651592, 6625427, 4818957, 4215156	7720510, 9168128, 10830341, 7196265	0.711	0.714	0.628

Table 3.7: Areas under the ROC curves for edgeR and baySeq for the three simulated datasets, while using the size factors estimated by sSeq (equivalently, by DESeq)

	Estimates $\hat{s}_{ij}$ by sSeq (and DESeq)			**AUROC**		
	Sim1	**Sim2**	**Sim3**	**Sim1**	**Sim2**	**Sim3**
edgeR	2674477×0.684, 6677525×1.729, 4832029×0.987, 4230815×0.866	2651592×0.686, 6625427×1.732, 4818957×0.984, 4215156×0.863	7720510×1.006, 9168128×1.196, 10830341×1.126, 7196265×0.749	0.853	0.862	0.846
baySeq	2674477×0.684, 6677525×1.729, 4832029×0.987, 4230815×0.866	2651592×0.686, 6625427×1.732, 4818957×0.984, 4215156×0.863	7720510×1.006, 9168128×1.196, 10830341×1.126, 7196265×0.749	0.848	0.858	0.900

Table 3.8: Areas under the ROC curves for edgeR and baySeq for the three simulated datasets, while using the true values of size factors used for the simulations

	True size factors s_{ij}			**AUROC**		
	Sim1	**Sim2**	**Sim3**	**Sim1**	**Sim2**	**Sim3**
sSeq	0.6, 1.5, 0.8, 0.7	0.6, 1.5, 0.8, 0.7	1.3, 1.6, 1.4, 0.9	0.974	0.977	0.896
edgeR	2674477×0.6, 6677525×1.5, 4832029×0.8, 4230815×0.7	2651592×0.6, 6625427×1.5, 4818957×0.8, 4215156×0.7	7720510×1.3, 9168128×1.6, 10830341×1.4, 7196265×0.9	0.912	0.916	0.872
DESeq	0.6, 1.5, 0.8, 0.7	0.6, 1.5, 0.8, 0.7	1.3, 1.6, 1.4, 0.9	0.967	0.971	0.869
baySeq	2674477×0.6, 6677525×1.5, 4832029×0.8, 4230815×0.7	2651592×0.6, 6625427×1.5, 4818957×0.8, 4215156×0.7	7720510×1.3, 9168128×1.6, 10830341×1.4, 7196265×0.9	0.890	0.899	0.923

sSeq can produce meaningful results in under-replicated RNA-seq screens. However we'd like to stress that RNA-seq screens do not eliminate the biological variation in gene expression [12]. As evidenced by Table 3.2 and Fig. 3.3, the under-replicated screens have lower reproducibility as compared to the replicated studies. Multiple biological replicates are necessary to adequately assess the full extent of the variation in the biological system. Therefore the under-replicated screens can only be conducted when followed by a rigorous experimental validation with complementary technologies and adequate sample size.

4. MS/MS WITH SWATH AQUISITION

4.1 Introduction

Proteins are macro-molecules folded in three dimensions. A protein consists of one or more peptides (i.e. chains of molecules that may chemically bind to the others). The sequence information of peptides and their abundance can be used for protein identification and quantification.

A well-known technique, mass spectrometry (MS), is typically used to produce the spectra of masses of the peptides in a sample, which are commonly named as MS1 spectra. However, information from these spectra is not specific enough to distinguish the peptides that have the same or similar masses and co-elute in the same time. Since different peptides consist of different fragments, the MS2 spectra of the masses of the peptides' fragments are required for peptide identification and quantification. This technique is Tandem mass spectrometry (MS/MS) that generates MS1 and MS2 spectra.

The methods used for target approach are Selected Reaction Monitoring (SRM) or Multiple Reaction Monitoring (MRM). They are well accepted methods that provide high specificity and sensitivity for protein quantification. When profiling samples in MS/MS instrumentation, specific peptides are selected. Prior knowledge (i.e. mass to charge ratio and retention time) of those peptides is provided. Only the MS2 spectra of those selected peptides' fragments are generated through the instrument no matter how many other peptides are presented in the sample. When additional peptides are selected, the samples have to be profiled again through the instrument. This problem substantially increases the experimental cost. Consequently, the Sequential Window Acquisition of All THeoretical fragment-ion spectra (SWATH) method is developed as a data-independent acquisition (DIA) method. It provides MS1 and MS2 spectra for all the peptides in a sample at once.

Instead of setting any prior knowledge for the instrument, the knowledge is utilized for the post identification and quantification of targeted peptides.

Unlike SRM or MRM that scan the fragments of only one peptide in a MS2 spectrum, SWATH produces a MS2 spectrum that includes the fragments of one or multiple peptides. Such experimental design substantially alleviates the workload of MS/MS instrumentation. This is a main reason why the SWATH acquisition method significantly reduces the experimental cost and shortens entire time period. In [5], it was illustrated that the SWATH method can also provide good sensitivity and specificity in identification compare SRM. However, it also brings challenges to the computational and statistical analysis of protein quantification because data is more noisy than before.

4.2 Background

Specifically in a run of the SWATH experiment, for example, there can be 32 windows for peptides (i.e. precursor ions) from 400-426 Da to 1175-1201 Da. At each step, a MS1 spectrum scans all the precursors. The MS2 spectra sequentially scan the fragments of the peptides present in each of the 32 windows. The 32 fragmentations take about 3.2 seconds to be completed. This procedure is repeated over thousands of times in a range of time period between 0 and 120 minutes.

A publicly available software [5], openSWATH, can be utilized to pick up the peaks of fragments based on eXtracted Ion Chromatograms (XICs). The purpose of using XICs is to reduce the three-dimensional data (i.e. count, retention time, and m/z) of MS2 spectra into the two-dimensional space (i.e. summed-up count and retention time). Practically, the spectrum of a fragment ion is often centered on a specific m/z value and varies within a 50 ppm bin equivalent to (theoretical mass)$\times 50 \times 10^{-6}$ in dalton. The counts of molecular ions that have m/z value within the 50 ppm bin are summed up to represent the abundance of this fragment. Such ions are extracted at each point in time. The XIC of the fragment consists of the summed-up counts across those ions. When a fragment is truly present in the sample, there should be at least a peak observed in its XIC.

After the intervals of retention time for all the observed peaks are estimated by openSWATH, another systematic software mProphet can be used to filter out those problematic peaks according to the pattern between the XICs of different fragments within the peptides. Combining openSWATH and mProphet, it enables the successful identification of meaningful peptides.

However, the previous methods just simply sums up the counts within the m/z bins at each point in time. They ignore the problem that the ions within those m/z bins may actually belong to different peptides but still generate large peak in the XIC over elution time. This may leave undetected interference noise into fragment quantification. In this project we argue that, in addition to the coherent pattern between XICs, it is challenging but necessary to also quantify the homogeneous pattern within XICs.

4.3 Methods

When the ions profiled in the same m/z bins belong to the same peptide, their traces over elution time share similar peak patterns. Motivated by this concept, we separate each XIC into several miniXICs in the m/z dimension. Those miniXICs are generated in tiny bins with a unit of 0.01 Da, and are utilized to account for the within-XIC information. We propose to fit a linear quadratic regression model in those miniXICs. The probability of having a negative quadratic coefficient is utilized to score the homogeneity of co-elution peaks. Details are provided as follows.

4.3.1 Per-fragment linear quadratic regression.

We denote $u_{lt,fpsg}$ as the count in log scale for the molecules in the l^{th} miniXIC at point t in retention time for fragment f of peptide p in sample s under group g. We fit the quadratic model Eq. (4.1) in all the miniXICs from fragment f.

$$\begin{gathered} u_{lt,fpsg} = \beta_{0,fpsg} + \beta_{1,fpsg} x_{t,psg} + \beta_{2,fpsg} x_{t,psg}^2 + \epsilon_{lt,fpsg}, \\ \text{for } l = 1, 2, \ldots, L_{fpsg},\ t = 1, 2, \ldots, T_{psg}, \text{ and } n_{fpsg} = L_{fpsg} \cdot T_{psg} \\ \text{with } \beta_{2,fpsg} \sim \mathcal{N}(\mu_{2,fpsg},\ \sigma^2_{2,fpsg}) \text{ and } \epsilon_{lt,fpsg} \sim \mathcal{N}(0,\ \sigma^2_{\epsilon,fpsg}) \end{gathered} \tag{4.1}$$

where $x_{t,psg}$, $\beta_{0,fpsg}$, $\beta_{1,fpsg}$, and $\beta_{2,fpsg}$ are the independent variable, the intercept, the linear coefficient, the quadratic coefficient. The characteristics of these model parameters are interpreted in the following paragraphs.

The independent variable $x_{t,psg}$ corresponds to the index of point in retention time. This value is the same for all the fragments within the same peptide. It should be noticed that different peptides can have different spans of retention time. The scale of the independent variable varies when fitting Eq. (4.1) in different fragments. As a result, the scale of the estimates of the model coefficients can be also different for different fragments. This makes the results not directly comparable between fragments. In order to avoid this problem, the values of the independent variable are standardizes as shown in Eq. (4.2).

$$\begin{gathered} x_{t,psg} = (t - \bar{t})/(T_{psg} - \bar{t}) \text{ and thus } x_{t,psg} \in [-1,\ 1] \\ \text{where } t = 1,\ 2,\ 3,\ .....,\ T_{psg} \text{ and } \bar{t} = \frac{1}{T_{psg}} \sum_{t=1}^{T_{psg}} t \end{gathered} \tag{4.2}$$

The intercept $\beta_{0,fpsg}$ represents the average height of peaks at the center. At each point in retention time, the intensities of a fragment in the m/z dimension typically present a bell shape centered at the true m/z value of that fragment. The miniXIC is a transection of the bell in the retention time dimension. Consequently, the peak height of a miniXIC can vary between the bottom and the largest intensity for the fragment. The intercept parameter

summarizes the common height across the miniXICs for fragment f of sample s in group g, and the variation of height between miniXICs involves the model error parameter $\epsilon_{lt,fpsg}$.

The linear coefficient $\beta_{1,fpsg}$ represents the location of apex in retention time. In the retention time dimension, the peaks of miniXIC are not often centered around the middle point in retention time. The linear coefficient is used to address this problem when the quadratic coefficient is fixed. This is illustrated in Eq. (4.3) that re-formularizes Eq. (4.1). The transformation indicates that the peak of XIC centers around $\beta_{1,fpsg}/(2\beta_{2,fpsg})$ which can be a non-zero value.

$$u_{lt,fpsg} = \beta_{2,fpsg}\left(x_{t,psg} + \beta_{1,fpsg}/(2\beta_{2,fpsg})\right)^2 + \epsilon_{lt,fpsg} + \text{constant}_{\cdot,fsg} \tag{4.3}$$
$$\text{where constant}_{\cdot,fsg} = \beta_{0,fpsg} - \beta^2_{1,fpsg}/(4\beta_{2,fpsg})$$

In Fig. 4.4-Fig. 4.3 under column 2, the observed miniXICs are the black lines. The fitted values using Eq. (4.1) are the read lines. In those graphs, we can see that the estimated peak hight is not necessarily on the point of time $x_{t,psg} = 0$. Furthermore, the position of peak apex can be quite different from the center of retention time.

The quadratic coefficient $\beta_{2,fpsg}$ represents the peak shape, and its sign does matter. In Eq. (4.3), it indicates that $u_{lt,fpsg}$ reaches the maximum at $x = -\beta_{1,fpsg}/(2\beta_{2,fpsg})$ when the peak shape parameter is negative, i.e. $\beta_{1,fpsg} < 0$. On the other side, $u_{lt,fpsg}$ changes to have a convex shape when $\beta_{1,fpsg} > 0$. The interference noise exists in such a situation. It is stronger when $\beta_{1,fpsg}$ is larger, and diminishes when $\beta_{1,fpsg}$ reaches negative infinity. Accordingly, we propose to calculate the probability of $\beta_{1,fpsg} < 0$ and utilize this probability to account for the potential effect due to the interference noise within a XIC.

Furthermore, in order to account for the random variation between miniXICs, we assume that the quadratic coefficient follows a Normal distribution with a mean $\mu_{2,fpsg}$ representing the common peak shape and a variance $\sigma^2_{2,fpsg}$ addressing variability among miniXICs. This random factor shrinks per-miniXIC regression estimates towards a stable target optimized through restricted maximum likelihood (REML) [99–101].

We propose a score w_{fpsg} that quantifies the strength of homogeneity in peaks of miniXICs. First, we describe the hypotheses in Eq. (4.4) and specify the statistic in Eq. (4.5) that follows a student t distribution under the null hypothesis H_0 with the degree of freedom as $df_{fpsg} = n_{fpsg} - 3$. The estimates of peak shape, $\hat{\mu}_{2,fpsg}$, and the standard error of those estimates are obtained using the open source R package `lme4` [101] with the function `lmer`. Finally, we calculate the score shown in Eq. (4.6) using the R package `stats` with the function `pt`.

$$H_0 : \ \mu_{2,fpsg} < 0 \quad \text{vs} \quad H_a : \ \mu_{2,fpsg} \geq 0 \tag{4.4}$$

$$\text{stat}_{fpsg} = \hat{\mu}_{2,fpsg}/\text{StandardError}(\hat{\mu}_{2,fpsg}) \tag{4.5}$$

$$w_{fpsg} = \text{Probability}(studentt_{df_{fpsg}} > \text{stat}_{fpsg} \text{ when } H_0 \text{ is TRUE}) \tag{4.6}$$

where $\text{StandardError}(\hat{\mu}_{2,fpsg}) = \sum_l \sum_t (u_{lt,fpsg} - \hat{u}_{lt,fpsg})^2 / \sum_t (x^2_{t,psg} - \overline{x^2}_{t,psg})^2$. The nominator in the formula of StandError is the total square of model residuals. It is affected by both the variance of the model error $\sigma^2_{\epsilon,fpsg}$ and the variance of the random factor $\sigma^2_{2,fpsg}$ in Eq. (4.1). The graphs in Fig. 4.4 are the fragments of a peptide that have very low weighting score, e.g between 0 and 0.004. Within the XICs of those fragments, the homogeneous patter between miniXICs (shown in column 1 and 2) is ambiguous. An example of perfect peak shape can be shown by Fig. 4.1 row 2. The weighting score is 1. Furthermore, Fig. 4.1 and Fig. 4.2 illustrate that a bell shape in peaks is not required for measuring the homogeneity within the XIC of a fragment.

4.3.2 All-fragment linear quadratic regression for a peptide.

As an extension to the per-fragment modeling, the model in Eq. (4.1) can be changed (shown in Eq. (4.7)) and fitted into the miniXICs across all the features within a peptide.

The purpose of using this extended all-fragment model is to share information among fragments.

$$\begin{aligned} u_{ltf,psg} &= \beta_{0f,psg} + \beta_{1,psg} x_{t,psg} + \beta_{2f,psg} x_{t,psg}^2 + \epsilon_{ltf,psg}, \\ &\text{for } l = 1, 2, \ldots, L_{fpsg},\ t = 1, 2, \ldots, T_{psg}, \\ &f = 1, 2, \ldots, F_{psg} \text{ and } n_{psg} = T_{psg} \sum_{f=1}^{F_{psg}} L_{fpsg} \\ \text{with } \beta_{2f,psg} &\sim \mathcal{N}(\mu_{2f,psg},\ \sigma_{2f,psg}^2) \text{ and } \epsilon_{ltf,psg} \sim \mathcal{N}(0,\ \sigma_{\epsilon,psg}^2) \end{aligned} \tag{4.7}$$

Since the heights of different fragments are different, the model intercept $\beta_{0f,psg}$ should be different upon f. For the linear coefficient, because the co-elution peaks of the fragments from the same peptide should have centers very similar to each other, we use $\beta_{1,psg}$ to account for the common position of apexes. Finally for the quadratic coefficient, we preserve the assumption in Eq. (4.1) and allow different variance components for different fragments. According to Eq. (4.8), we calculate the score that reflects the within-XIC variation based on the all-fragment model in Eq. (4.7). The degree of freedom in this model is $df_{psg} = (n_{psg} - 1) - 2 \times F_{psg}$.

$$\begin{aligned} \text{stat}_{f,psg} &= \hat{\mu}_{2f,psg} / \text{StandError}(\hat{\mu}_{2f,psg}) \\ w_{f,psg} &= \text{Probability}(studentt_{df_{psg}} > \text{stat}_{f,psg} \text{ when } H_0 \text{ is TRUE}) \end{aligned} \tag{4.8}$$

4.3.3 Evaluations

Fitting the linear fixed-effect model into the raw intensity. For each protein, Eq. (4.9) is developed in [102] and utilized to detect any differential changes.

$$\begin{aligned} y_{fpsg} = \ & \mu_{\cdot} + \mathcal{F}_f + S(G)_{s(g)} + G_g + \varepsilon_{fsg} \\ & \text{with } \varepsilon_{fpsg} \sim \mathcal{N}(0,\ \sigma^2) \end{aligned} \tag{4.9}$$

where $\mu_{\cdot}$ is the average intensity. $\mathcal{F}_f$ is the statistical fixed effect on the raw intensity y_{fpsg} due to fragment ion f in the protein. $S(G)_{s(g)}$ and G_g is the statistical effects due to subject s and group g. This model is implemented in the open source R package `SRMstats` [102].

The raw intensity can be either the total count under the XIC curve or the peak height of the curve. The performance is illustrated in Fig. 4.6. Specifically, the term logCnt in the figure is the total count under the curve in log scale, that is $y_{fpsg} = \sum_l \sum_t u_{lt,fpsg}$. On the other hand, the intensity can also be the estimated peak height of miniXICs, i.e. $y_{fpsg} = \hat{\beta}_{0,fpsg}$ in the per-fragment model Eq. (4.1) or $y_{f,psg} = \hat{\beta}_{0,sg}$ in the all-fragment model Eq. (4.7).

Fitting the linear fixed-effect model with unequal variance adjusted by the weighting score. After obtaining the probability score w_{fpsg} for each fragment, we evaluate the performance of the proposed score by comparing the results of protein differential analysis using the weight or not. In the study of protein differential analysis, a linear fixed-effect model for each protein shown in Eq. (4.9) is adjusted in Eq. (4.10). In order to weight the raw intensity, we assume the model has unequal variance and adjust this heterogeneity by the proposed weighting score w_{fpsg}. In Eq. (4.10), the parameters and variables are denoted as the same as those in Eq. (4.9) except for that the model error has unequal variance between fragments.

$$\begin{aligned} y^*_{fpsg} = \quad & \mu^*_{\cdot} + \mathcal{F}^*_f + S(G)^*_{s(g)} + G^*_g + \varepsilon^*_{fpsg} \\ & \text{with } \varepsilon^*_{fpsg} \sim \mathcal{N}(0,\ w^2_{fpsg} \cdot \sigma^2) \end{aligned} \tag{4.10}$$

For the purpose of comparison, we use an alternative method, mScore, which is originally provided by the open source software mProphet [103]. It gives one value for one peptide within a sample, and accounts for the between-XIC coherence across fragments. The original purpose of using mScore is to control the false discovery rate. It linearly combines several scores the quantify the coherent pattern between fragments within the peptide. The smaller the value is, the stronger the pattern is. In order to make this alternative approach comparable to the proposed approach, we convert it into standardized value in Eq. (4.11).

$$\text{mScore} = -log(\text{m\_score})/max(-log(\text{m\_score})) \tag{4.11}$$

where m_score is the original value provided by the software. The standardized score has values between 0 and 1. When mScore is greater, the between-fragment coherency is

stronger. In order to adjust the raw intensity by mScore, the model assumption in Eq. (4.10) is changed to be $\varepsilon^*_{fpsg} \sim \mathcal{N}(0,\ \text{mScore}^2_{psg} \cdot \sigma^2)$.

4.3.4 Dataset

A SWATH-MS experiment of yeast organism [5] involves two conditions, i.e. 0 minutes (A) and 120 minutes (B). Three biological complex samples are processed under each of the two conditions. There should be thousands of proteins in one complex sample, e.g. around 2200$\sim$2340. The size of the raw data is large, such as 10$\sim$22GB for just one run. Before performing differential analysis for proteins, peptides and their fragments are identified by Spectronaut and filtered by m_score. All the peptides that have m_score greater than 0.01 are considered as false identified. As a result, only the peptides that have m_score less than 0.01 are used for evaluation. Since the majority of the peptides are true, more than 50$\%$ of them obtain the proposed weighting score w_{fpsg} greater than 0.9.

The number of differentially changed proteins between A and B is expected to be larger than the number of changed proteins between samples within condition A. The result of comparison between A and B empirically reflects the sensitivity (i.e. AB), and the result of comparison between samples within condition A can reflect the specificity (i.e.AA). ECDF plots of p-values are utilized to visualize the efficacy of the proposed score. Such an evaluation method has been used in [4].

4.4 Results

the ECDF curve of p-values for between-condition comparisons (AB) empirically indicates the sensitivity. When it is closer to the top left corner, the method is more sensitive. The ECDF curve of p-values for within-condition comparisons (AA) empirically indicates the specificity. When it is closer to the 45 degree line, the method has higher specificity and smaller false discovery rate.

The proposed score improves the result of protein differential analysis. In Fig. 4.6 (a)-(d), adjusting with the proposed score (termed as w) always increases both the sensitiv-

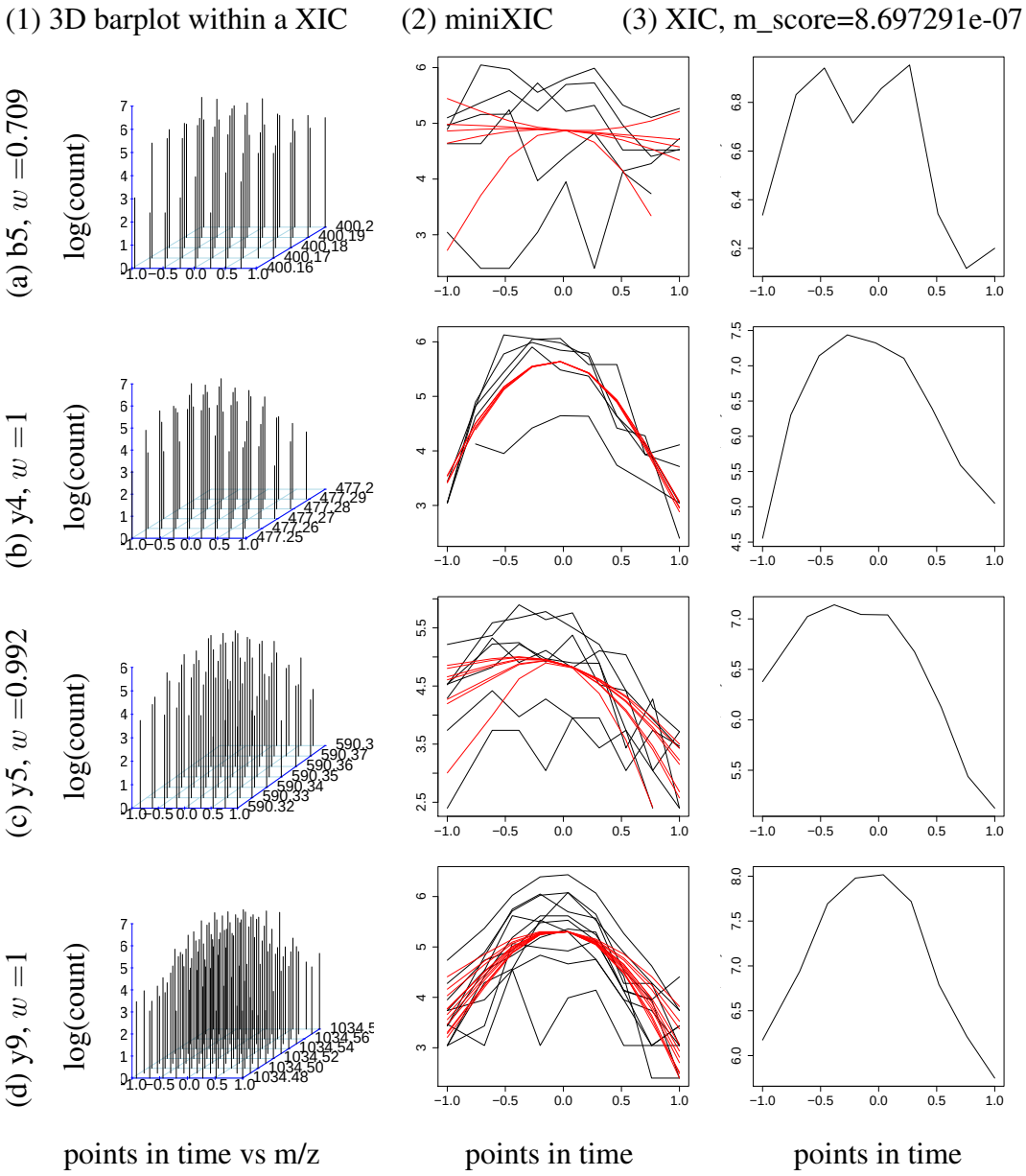


Fig. 4.1.: The proposed score is close to 1 when miniXICs of the fragment share homogeneous peak shape, illustrated by peptide AAADALSDLEIKDSK in sample s=1 and group g=1. Column 1 is the three-dimensional barplots of intensities. Column 2 overlays the miniXICs (black lines) and the fitted lines (red lines) with Eq. (4.1). The XIC plot of a fragment shown in column 3 is the total intensity at each point in time based on the 3D barplot shown in column 1. The X-axis in all the graph is the independent variable x in Eq. (4.1) and Eq. (4.2).

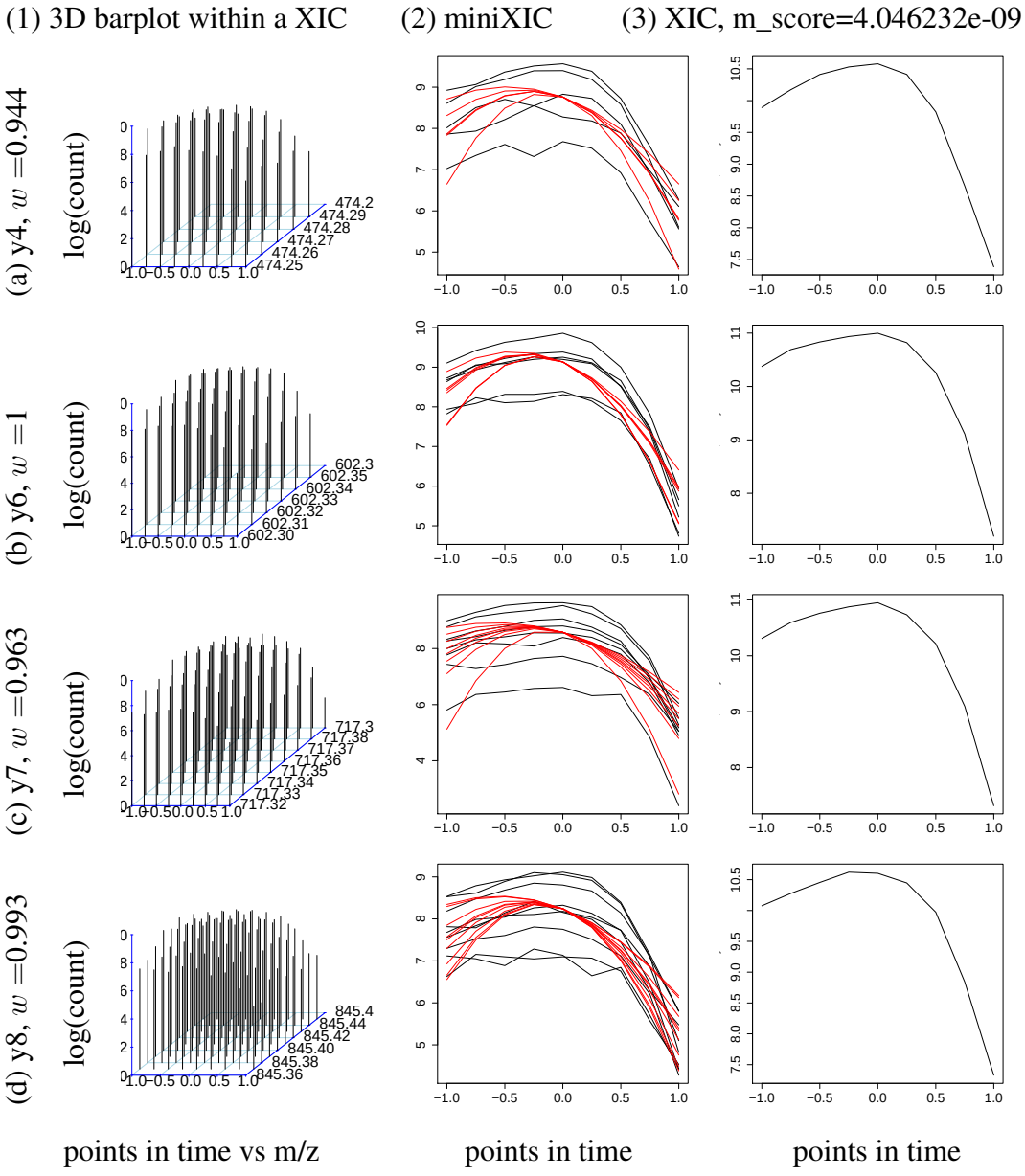


Fig. 4.2.: The proposed score gives high weight to the fragment as long as an apex observed in only a partial peak, illustrated by peptide YAQDGAGIER in sample s=6 and group g=2. Axes and labels are as in Fig. 4.1.

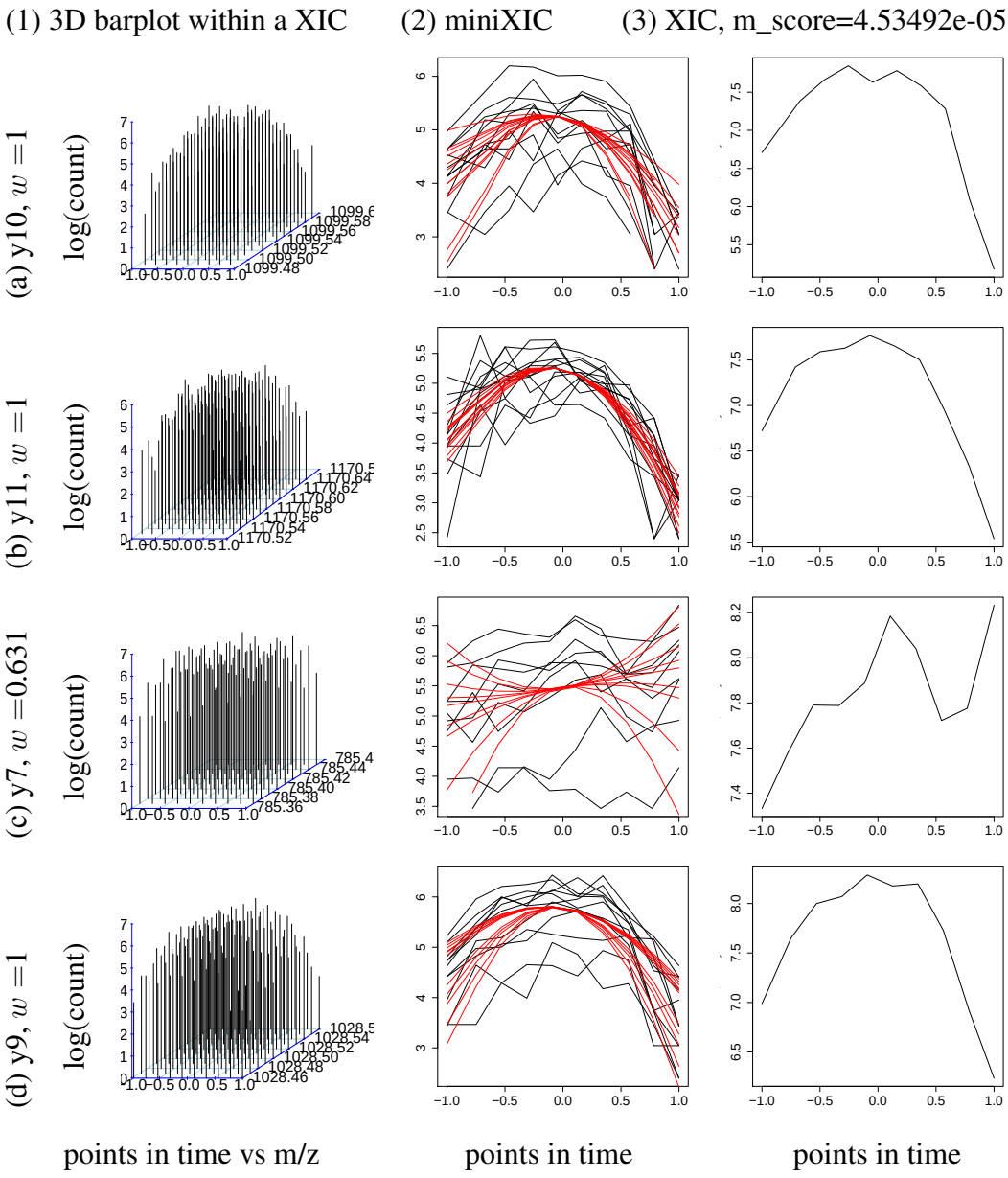


Fig. 4.3.: The proposed score gives low weight to the fragment when its miniXICs have flat pattern independently from the other fragments within the peptide AAQDSF AAGWGVMVSHR in sample s=2 and group g=1. Axes and labels are as in Fig. 4.1. The fragment with interference noise is illustrated in row (c).

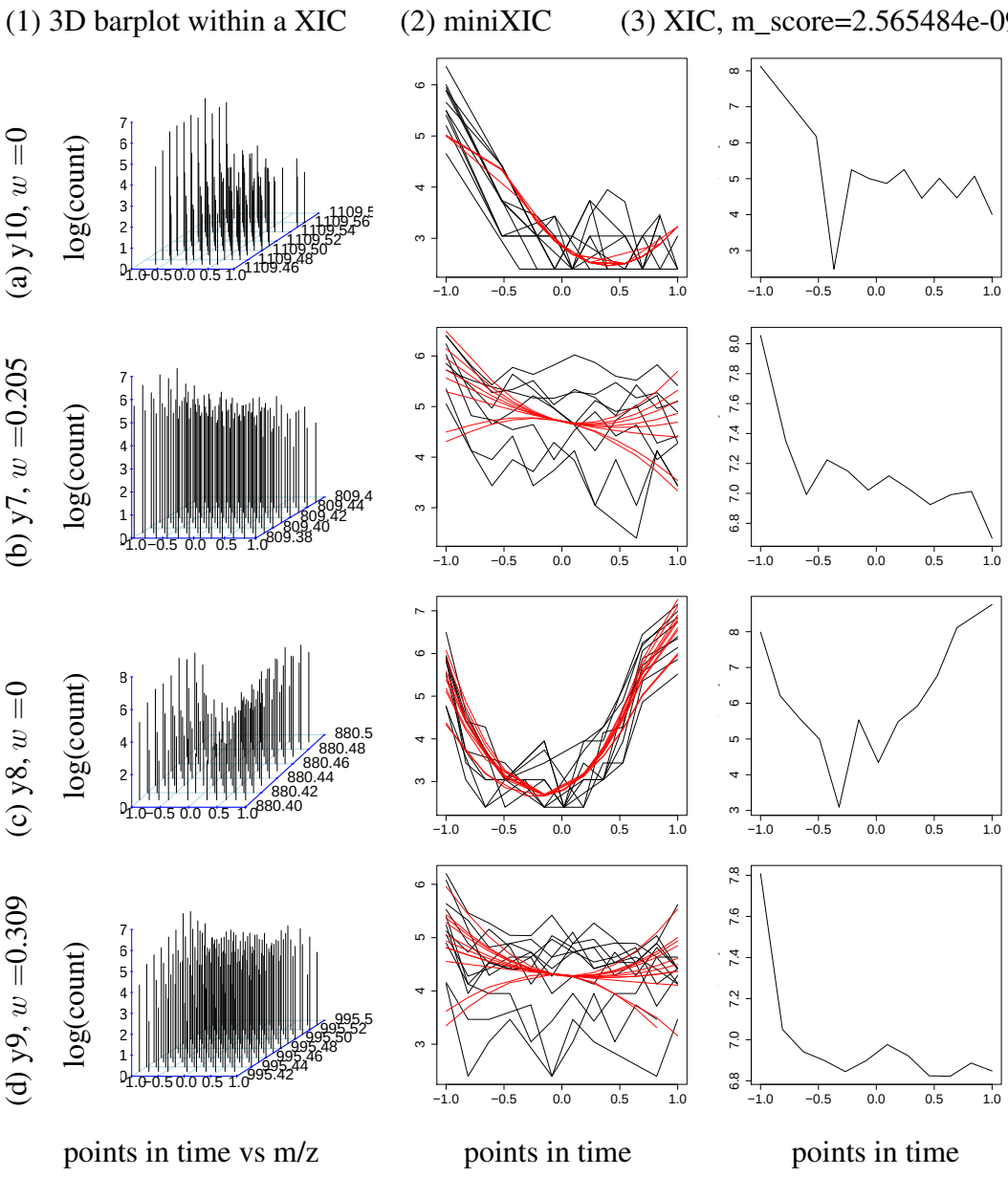


Fig. 4.4.: The proposed score is close to 0 when the interference noise is strong, illustrated by peptide SKLNDAVEYVSGR2 in a sample. Axes are as in Fig. 4.1.

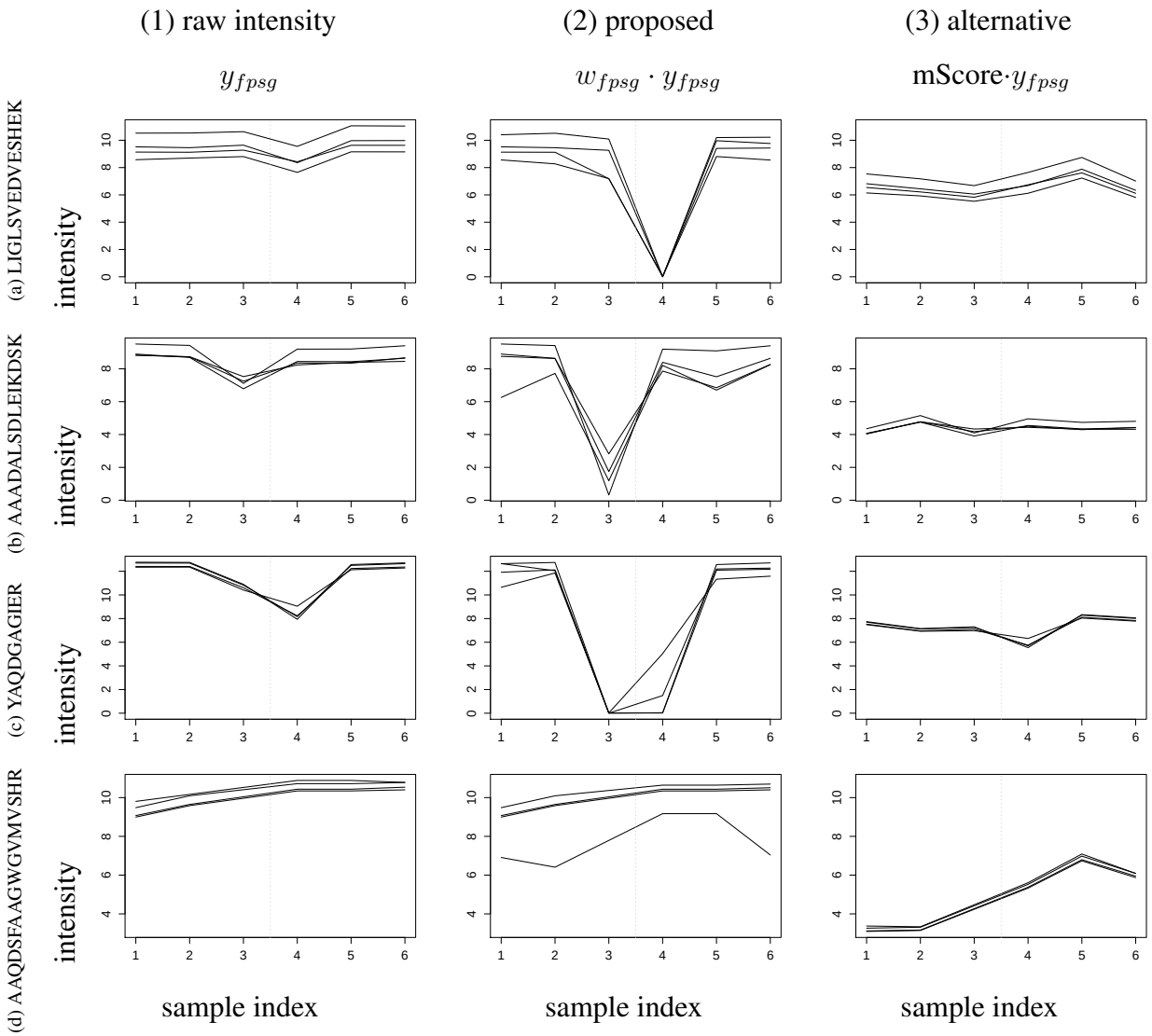


Fig. 4.5.: The proposed method succeeds in providing consistent weighting scores within a sample or a fragment. The profile plots shown in this figure are fragments' intensity versus sample index. A connected line in the plot is the profile of a fragment. The first 3 samples are under condition A and the last 3 samples are under condition B. This is illustrated by the profile plots of the four peptides shown in Fig. 4.4-Fig. 4.3. The four peptides shown in Fig. 4.4-Fig. 4.3 are presented in rows of this figure. It is shown that the weighting effect of mScore is ignorable. Detailed information about the formulae and the interpretation are provided in Chapter 4.3 and Chapter 4.3.3. The Y-axis in column 1 is the total raw intensity under the XIC curve of a fragment. It is the dependent variable in Eq. (4.9). The Y-axis in column 2 and 3 is the dependent variable in Eq. (4.10).

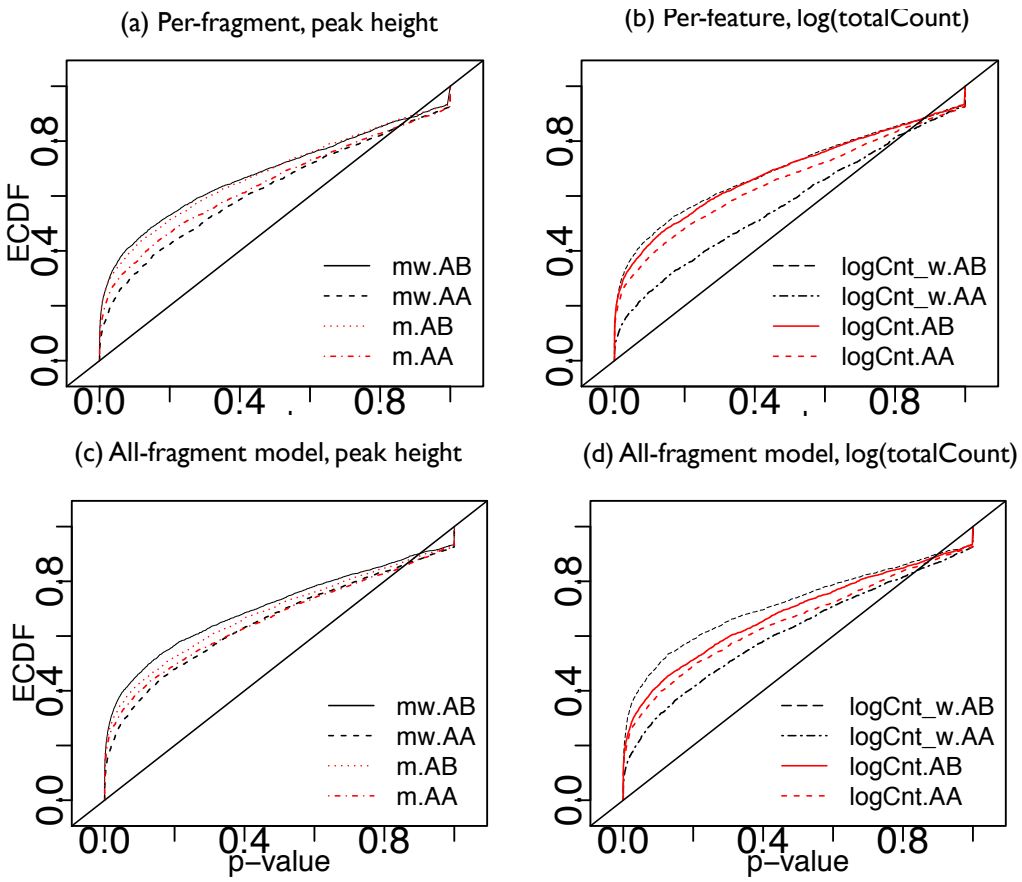


Fig. 4.6.: The total intensity adjusted by the proposed weight score improves the accuracy of testing results upon this dataset. `AB` and `AA` label the comparison between conditions and the comparison within conditions. `m.AB` and `m.AA` are the quantification with estimated peak heights. `mw.AB` and `mw.AA` are the products between peak heights and the weighting score. `logCnt.AB` and `logCnt.AA` are the quantification with log of total count. `logCnt_w.AB` and `logCnt_w.AA` are the products between logCnt and the weighting score (see Chapter 4.3.3 and Chapter 4.4 for details).

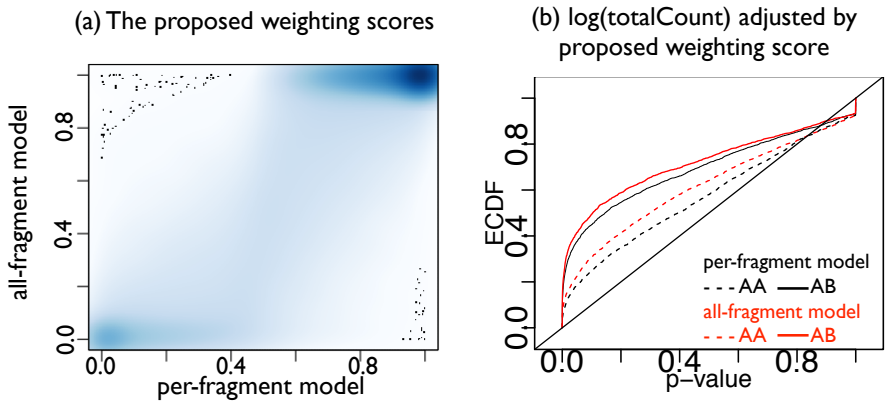


Fig. 4.7.: The comparison between per-fragment models and all-fragment models indicates that they are competitive. (a) The smooth scatter plot of all-fragment model vs per-fragment model indicates that the most of the scores are similar since the majority is around the 40 degree line. The dark colored area at the top right indicates that more fragments obtained the weight close to 1 using all-fragment model than using per-fragment model. (b) The ECDF plot indicates that the proposed score using per-fragment model helps to increase specificity in tests and moderately sacrifice sensitivity compare all-fragment model.

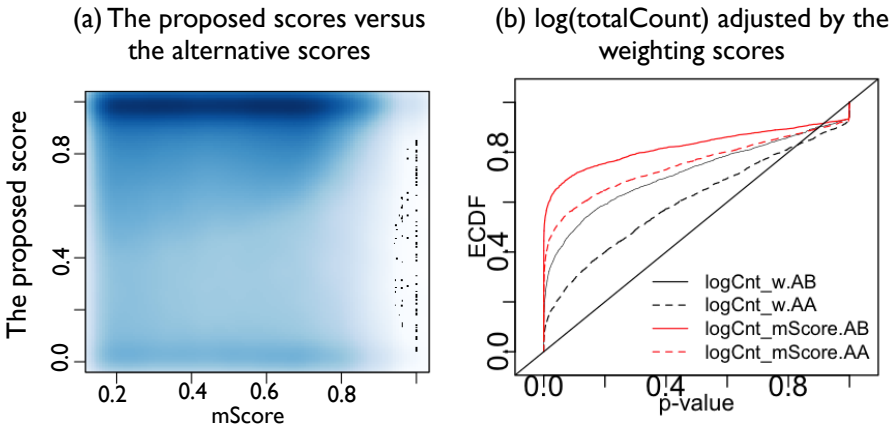


Fig. 4.8.: The proposed score outperforms the alternative score by increasing specificity in tests. (a) In this smooth scatter plot, mScore is standardized to be comparable with the proposed score. Details are provided in Chapter 4.3.3 The data shown in this plot is already filtered so that all the fragments have strong coherent pattern within the peptide. The purpose is to show that the proposed score is complementary to the existing alternative. It further improves the accuracy in test based on the filtration result from `mProphet`. (b) In the ECDF plot, axes and labels are as in Fig. 4.6. Per-fragment model is used for the proposed score. The ECDF curve of p-values for the within-condition comparison using the proposed score is much lower than the curve using the alternative. It indicates the reduction of false discovery rate.

ity and specificity. For example, `mw.AB` always has a higher curve than `m.AB`. And `mwAA` always has a lower curve than `m.AA`. The same result can be observed for `logCnt` and `logCnt_w`. This result illustrates that adjusting fragments' intensity with the proposed weighting score sufficiently powerful to increase the accuracy for the differential analysis of proteins.

The total intensity is better than the peak height for quantification. The total intensity is the summed-up counts under the XIC peak. This quantification method is currently used by the open source software `Spectronaut`. In Fig. 4.6, it is also observed that using the weighted peak heights to quantify fragments' abundance (i.e. `mw.AB` and `mAA`) can obtain similar sensitivity but much less specificity than the quantification based on the total intensity (i.e. `logCnt_w.AA` and `logCnt_w.AB`. Therefore, from now on, we further evaluate the proposed score by focusing on the quantification based on the total intensity in log scale.

Per-fragment models and all-fragment models are competitive. For the purpose of borrowing common information across fragments with the peptide, all-fragment models are proposed. The ECDF plot in Fig. 4.7 illustrates that the all-fragment models return higher intensity but slightly lower specificity. Howver, the difference is not as significant as the difference from the other approaches shown in Fig. 4.6.

The proposed score outperforms the alternative. As an existing weighting method, mScore (provided by the open source software `mProphet` is used to adjust the intensity of fragments. The result using this existing score is exemplified in Fig. 4.5 (column 3) and evaluated in Fig. 4.8. While increasing the sensitivity `logCnt_mScore.AB` in Fig. 4.8, the specificity `logCnt_mScore.AB` is substantially decreased. However, the result using the proposed score indicates a comparable sensitivity as well as much higher specificity than mScore.

4.5 Discussion

In the SWATH experiments, MS1 and MS2 spectra are obtained using a data-independent approach through instrumentation. They are then analyzed using a target approach for identification and quantification. Currently existing methods provide identification and quantification analysis based on the extracted ion chromatography (XIC) within the m/z bins. The variabilty between XICs of different fragments is addressed. However, the interference noise within a XIC is not yet accounted for. We propose a score that quantifies the strength of the homogeneous peak pattern among miniXICs within a XIC. This score is utilized as a weight. The intensity of a fragment obtained from its XIC is multiplied by the proposed score. The evaluation result indicates that both the empirical sensitivity and specificity of protein differential analysis are substantially improved. It also illustrates that the proposed score outperforms the alternative such as mScore provided in the open source software mProphet.

When calculating the proposed score, two models are proposed with different approaches. Per-fragment modeling emphasizes that the within-XICs variation is independent between different fragments. The homogeneity within a XIC is totally un-related to that within another XIC even though the two fragments may come from the same peptide. In order to avoid any potential mis-communication between genuine fragments and artificially identified fragments, per-fragment model is recommended.

Another approach is all-fragment modeling. It assumes that the identified fragments for a peptide are jointly genuine or jointly artificial. If the majority of the fragments present strong interference noise, the other fragments from that peptide are also weighted down. When all the identified fragments are truly from the same peptide, all-fragment modeling is sufficiently powerful to rescue the low-abundant fragments. It is illustrated in Chapter 4.4 that this approach increases the empirical sensitivity and moderately sacrifices the empirical specificity.

Furthermore, a discussion may involve the appropriateness of the fixed-effect modeling used for evaluation. The model in Eq. (4.9) (implemented with the open source R package

`SRMstats`) assumes that the variance component of model error ε_{fsg} is common across fragments. However, this assumption may not always be satisfied. Especially when the statistical interaction between fragment and model error is significant, a model with the assumption of unequal variance components may be required. However, the diagnosis of residuals needs to be performed before accounting for the unequal variance in a new model. For example, the presence of un-equal variance can be indicated by a strong association between absolute model residuals and the factor of fragments.

Finally, this chapter proposes a weighting score to adjust fragments' intensities in order to account for the within-XIC variation. Although the proposed approach increases the empirical sensitivity and specificity, the accuracy of protein differential analysis can be further improved by removing artificially identified fragments from the model fitting. The approach of feature selection with machine learning or data mining methods will be further developed in future research.

5. SUMMARY AND FUTURE RESEARH

In this thesis, we focus on the problem of estimation of variation that involves three types of important high-throughput biological molecular experiments. Because different experimental designs generate datasets with different characteristics, we propose different statistical methods to account for sources of variation. For the purpose of reducing noise in perturbation experiments in Chapter 2, we utilize the information from controls samples, and propose to stepwise estimate the additive and non-additive effects with linear mixed-effect models. In order to reduce the variation of variance estimates in the Negative Binomial models in Chapter 3, we propose to shrink the dispersion estimates towards a common information borrowed across genes. Finally in Chapter 4, working directly with large raw data (i.e. 10$\sim$25 GB), we propose a score that quantifies the strength of homogeneity within signals identified for a fragment. The purpose of the proposed method is to weigh down the fragment identified by `openSWATH` that actually consists of the ions generated from different peptides.

In future research in perturbation screens, gene-gene interaction can be identified from the analysis of pairs of gene perturbation instead of single gene perturbation. The number of mutant samples is much greater. The sources of additive and non-additive variance are more complicated. Further research with new statistical modeling is required.

RNA-seq experiments are considered as a complementary or even substitute method of microarray experiments for gene expression analysis. The tools and methods of processing and analyzing RNA-seq experiments are developing rapidly. Instead of developing methods for RNA-seq itself, a challenging but promising approach can be combining it with the other meaningful experiments, such as ChIP-seq experiments.

In the analysis of SWATH experiments, future research can endeavor to implement new models with unequal variance model error in the step of protein differential analysis. This new model needs to be applied especially when diagnosis of model residuals indicates the

significant dependence between the absolute residuals and the factor of fragments. Furthermore, the feature selection or machine learning methods (e.g. LASSO, ridge regression, or elastic net, etc) can be applied to filter out those misinterpreted fragments.

LIST OF REFERENCES

LIST OF REFERENCES

[1] F. Crick *et al.*, "Central dogma of molecular biology," *Nature*, vol. 227, no. 5258, pp. 561–563, 1970.

[2] D. Yu, J. Danku, I. Baxter, S. Kim, O. K. Vatamaniuk, D. E. Salt, and O. Vitek, "Noise reduction in genome-wide perturbation screens using linear mixed-effect models," *Bioinformatics*, vol. 27, no. 16, pp. 2173–2180, 2011.

[3] D. Yu, J. M. Danku, I. Baxter, S. Kim, O. K. Vatamaniuk, O. Vitek, M. Ouzzani, and D. E. Salt, "High-resolution genome-wide scan of genes, gene-networks and cellular systems impacting the yeast ionome," *BMC genomics*, vol. 13, no. 1, p. 623, 2012.

[4] D. Yu, W. Huber, and O. Vitek, "Shrinkage estimation of dispersion in negative binomial models for rna-seq experiments with small sample size," *Bioinformatics*, vol. 29, no. 10, pp. 1275–1282, 2013.

[5] L. C. Gillet, P. Navarro, S. Tate, H. Röst, N. Selevsek, L. Reiter, R. Bonner, and R. Aebersold, "Targeted data extraction of the ms/ms spectra generated by data-independent acquisition: a new concept for consistent and accurate proteome analysis," *Molecular & Cellular Proteomics*, vol. 11, no. 6, 2012.

[6] C. Boone, H. Bussey, and B. J. Andrews, "Exploring genetic interactions and networks with yeast," *Nat Rev Genet*, vol. 8, p. 437, 2007.

[7] M. Gstaiger and R. Aebersold, "Applying mass spectrometry-based proteomics to genetics, genomics and network biology.," *Nat Rev Genet*, vol. 10, pp. 617–627, Sep 2009.

[8] T. Ideker and R. Sharan, "Protein networks in disease," *Genome Res*, vol. 18, pp. 644–652, 2008.

[9] N. Bharucha and A. Kumar, "Yeast genomics and drug target identification.," *Comb Chem High Throughput Screen*, vol. 10, pp. 618–634, Sep 2007.

[10] M. Boutros and J. Ahringer, "The art and design of genetic screens: RNA interference.," *Nature reviews. Genetics*, vol. 9, pp. 554–66, 2008.

[11] S. L. Forsburg, "The art and design of genetic screens: yeast.," *Nat Rev Genet*, vol. 2, pp. 659–668, Sep 2001.

[12] F. Markowetz and R. Spang, "Inferring cellular networks – a review," *BMC Bioinformatics*, vol. 8, pp. 1–17, 2007. Review on inferring interactions.

[13] F. Markowetz, "How to understand the cell by breaking it: network analysis of gene perturbation screens," *PLoS Computational Biology*, 2010.

[14] X. D. Zhang and J. F. Heyse, "Determination of sample size in genome-scale rnai screens.," *Bioinformatics*, vol. 25, pp. 841–844, Apr 2009.

[15] N. Malo, J. A. Hanley, S. Cerquozzi, J. Pelletier, and R. Nadon, "Statistical practice in high-throughput screening data analysis," *Nature Biotechnology*, vol. 24, pp. 167–175, 2006.

[16] A. M. Wiles, D. Ravi, S. Bhavani, and A. J. R. Bishop, "An analysis of normalization methods for Drosophila RNAi genomic screens and development of a robust validation scheme.," *J Biomol Screen*, vol. 13, pp. 777–784, Sep 2008.

[17] J. Tukey, "A survey of sampling from contaminated distributions," *I. Olkin*, 1960.

[18] S. Collins, M. Schuldiner, N. Krogan, and J. Weissman, "A strategy for extracting and analyzing large-scale quantitative epistatic interaction data," *Genome biology*, vol. 7, no. 7, p. R63, 2006.

[19] Y. H. Yang, S. Dudoit, P. Luu, D. M. Lin, V. Peng, J. Ngai, and T. P. Speed, "Normalization for cdna microarray data: a robust composite method addressing single and multiple slide systematic variation.," *Nucleic Acids Res*, vol. 30, p. e15, Feb 2002.

[20] B. M. Bolstad, R. A. Irizarry, M. Astrand, and T. P. Speed, "A comparison of normalization methods for high density oligonucleotide array data based on variance and bias.," *Bioinformatics*, vol. 19, pp. 185–193, Jan 2003.

[21] A. Bankhead, I. Sach, C. Ni, N. LeMeur, M. Kruger, M. Ferrer, R. Gentleman, and C. Rohl, "Knowledge based identification of essential signaling from genome-scale siRNA experiments," *BMC Systems Biology*, vol. 3, p. 80, 2009.

[22] J. Leek, R. Scharpf, H. Bravo, D. Simcha, B. Langmead, W. Johnson, D. Geman, K. Baggerly, and R. Irizarry, "Tackling the widespread and critical impact of batch effects in high-throughput data," *Nature Reviews Genetics*, vol. 11, no. 10, pp. 733–739, 2010.

[23] J. Leek and J. Storey, "Capturing heterogeneity in gene expression studies by surrogate variable analysis," *PLoS Genetics*, vol. 3, no. 9, p. e161, 2007.

[24] A. Birmingham, L. M. Selfors, T. Forster, D. Wrobel, C. J. Kennedy, E. Shanks, J. Santoyo-Lopez, D. J. Dunican, A. Long, D. Kelleher, Q. Smith, R. L. Beijersbergen, P. Ghazal, and C. E. Shamu, "Statistical methods for analysis of high-throughput rna interference screens.," *Nat Methods*, vol. 6, pp. 569–575, Aug 2009.

[25] N. Rieber, B. Knapp, R. Eils, and L. Kaderali, "RNAither, an automated pipeline for the statistical analysis of high-throughput RNAi screens.," *Bioinformatics (Oxford, England)*, vol. 25, pp. 678–9, March 2009.

[26] M. Boutros, L. P. Brás, and W. Huber, "Analysis of cell-based RNAi screens," *Genome Biology*, 2006.

[27] G. K. Smyth, "Linear models and empirical Bayes methods for assessing differential expression in microarray experiments," *Statistical Applications in Genetics and Molecular Biology*, vol. 3, 2004.

[28] G. K. Smyth, *Bioinformatics and Computational Biology Solutions Using R and Bioconductor*, ch. *Limma: Linear Models for Microarray Data*. Springer, 2005.

[29] Y. Benjamini and Y. Hochberg, "Controlling the false discovery rate: a practical and powerful approach to multiple testing," *Journal of the Royal Statistical Society*, vol. 57, pp. 289–300, 1995.

[30] B. Efron, "Microarrays, Empirical Bayes, and the two-groups model," *Statistical Science*, vol. 23, pp. 1–22, 2008.

[31] X. D. Zhang, P. F. Kuan, M. Ferrer, X. Shu, Y. C. Liu, A. T. Gates, P. Kunapuli, E. M. Stec, M. Xu, S. D. Marine, D. J. Holder, B. Strulovici, J. F. Heyse, and A. S. Espeseth, "Hit selection with false discovery rate control in genome-scale RNAi screens.," *Nucleic acids research*, vol. 36, pp. 4667–79, 2008.

[32] I. M. Kaplow, R. Singh, A. Friedman, C. Bakal, N. Perrimon, and B. Berger, "Rnaicut: automated detection of significant genes from functional genomic screens.," *Nat Methods*, vol. 6, pp. 476–477, Jul 2009.

[33] R. Wolfinger, G. Gibson, E. Wolfinger, L. Bennett, H. Hamadeh, P. Bushel, C. Afshari, and R. Paules, "Assessing gene significance from cDNA microarray expression data via mixed models," *Journal of Computational Biology*, vol. 8, pp. 625–637, 2001.

[34] K. Dobbin and R. Simon, "Comparison of microarray designs for class comparison and class discovery," *Bioinformatics*, vol. 18, p. 1438, 2002.

[35] M. Lindstrom and D. Bates, "Newton-Raphson and EM Algorithms for Linear Mixed-Effects Models for Repeated-Measures Data," *Journal of the American Statistical Association*, vol. 83, pp. 1014–1022, 1988.

[36] A. Baryshnikova, M. Costanzo, Y. Kim, H. Ding, J. Koh, K. Toufighi, J. Youn, J. Ou, B. San Luis, S. Bandyopadhyay, *et al.*, "Quantitative analysis of fitness and genetic interactions in yeast on a genome scale," *Nature Methods*, vol. 7, no. 12, pp. 1017–1024, 2010.

[37] D. Hoaglin, F. Mosteller, and J. Tukey, *Understanding robust and exploratory data analysis*. John Wiley & Sons. pp. 404–414, 1983.

[38] D. E. Salt, I. Baxter, and B. Lahner, "Ionomics and the study of the plant ionome.," *Annu Rev Plant Biol*, vol. 59, pp. 709–733, 2008.

[39] I. Baxter, "Ionomics: studying the social network of mineral nutrients.," *Curr Opin Plant Biol*, vol. 12, pp. 381–386, Jun 2009.

[40] J. Danku, L. Gumaelius, I. Baxter, and D. Salt, "A high-throughput method for Saccharomyces cerevisiae (yeast) ionomics," *Journal of Analytical Atomic Spectrometry*, vol. 24, pp. 103–107, 2009.

[41] D. J. Eide, S. Clark, T. M. Nair, M. Gehl, M. Gribskov, M. L. Guerinot, and J. F. Harper, "Characterization of the yeast ionome: a genome-wide analysis of nutrient mineral and trace element homeostasis in saccharomyces cerevisiae.," *Genome Biol*, vol. 6, p. R77, 2005.

[42] X. F. Liu, F. Supek, N. Nelson, and V. C. Culotta, "Negative control of heavy metal uptake by the Saccharomyces cerevisiae BSD2 gene," *Journal of Biological Chemistry*, vol. 272, pp. 11763–11769, 1997.

[43] C. M. Lauer Jùnior, D. Bonatto, A. A. Mielniczki-Pereira, A. Z. Schuch, J. F. Dias, M. L. Yoneama, and J. A. P. Henriques, "The PMR1 protein, the major yeast Ca2+-ATPase in the Golgi, regulates intracellular levels of the cadmium ion," *FEMS Microbiol Letters*, vol. 285, pp. 79–88, 2008.

[44] H. K. Rudolph, A. Antebi, G. R. Fink, C. M. Buckley, T. E. Dorman, J.LeVitre, L. S. Davidow, J. I. Mao, and D. T. Moir, "The yeast secretory pathway is perturbed by mutations in PMR1, a member of a Ca2+ ATPase family," *Cell*, vol. 58, pp. 133–145, 11989.

[45] M. Aouida, A. Khodami-Pour, and D. Ramotar, "Novel role for the Saccharomyces cerevisiae oligopeptide transporter Opt2 in drug detoxification," *Biochem Cell Biol*, vol. 87, pp. 653–661, 2009.

[46] C. Stark, B.-J. Breitkreutz, T. Reguly, L. Boucher, A. Breitkreutz, and M. Tyers, "Biogrid: a general repository for interaction datasets," *Nucleic acids research*, vol. 34, no. suppl 1, pp. D535–D539, 2006.

[47] Z. Wang *et al.*, "RNA-seq: A revolutionary tool for transcriptomics," *Nature Reviews Genetics*, vol. 10, pp. 57–63, 2009.

[48] M. Metzker, "Sequencing technologies: The next generation," *Nature Reviews Genetics*, vol. 11, pp. 31–46, 2009.

[49] E. Mardis, "Next-generation DNA sequencing methods," *Annual Reviews in Genomics and Human Genetics*, vol. 9, pp. 387–402, 2008.

[50] M. Garber *et al.*, "Computational methods for transcriptome annotation and quantification using RNA-seq," *Nature Methods*, vol. 8, pp. 469–477, 2011.

[51] A. Oshlack *et al.*, "From RNA-seq reads to differential expression results," *Genome Biol.*, vol. 11, p. 220, 2010.

[52] S. Pepke *et al.*, "Computation for ChIP-seq and RNA-seq studies," *Nature Methods*, vol. 6, pp. S22–S32, 2009.

[53] S. Anders and W. Huber, "Differential expression analysis for sequence count data," *Genome Biol.*, vol. 11, p. R106, 2010.

[54] M. Robinson *et al.*, "EdgeR: A Bioconductor package for differential expression analysis of digital gene expression data," *Bioinformatics*, vol. 26, p. 139, 2010.

[55] M. Robinson and G. Smyth, "Moderated statistical tests for assessing differences in tag abundance," *Bioinformatics*, vol. 23, pp. 2881–2887, 2007.

[56] T. Hardcastle and K. Kelly, "BaySeq: Empirical Bayesian methods for identifying differential expression in sequence count data," *BMC Bioinformatics*, vol. 11, p. 422, 2010.

[57] J. Li and R. Tibshirani, "Finding consistent patterns: A nonparametric approach for identifying differential expression in RNA-seq data," *To appear, Stat. Methods in Medical Research*, 2012.

[58] Y. Zhou *et al.*, "A powerful and flexible approach to the analysis of RNA sequence count data," *Bioinformatics*, vol. 27, pp. 2672–2678, 2011.

[59] P. Auer and R. Doerge, "A two-stage Poisson model for testing RNA-seq data," *Statistical Applications in Genetics and Mol. Biol.*, vol. 10, pp. 1–26, 2011.

[60] J. Li *et al.*, "Normalization, testing, and false discovery rate estimation for RNA-sequencing data," *Biostatistics*, vol. 13, pp. 523–538, 2011.

[61] J. Marioni *et al.*, "RNA-seq: An assessment of technical reproducibility and comparison with gene expression arrays," *Genome Res.*, vol. 18, pp. 1509–1517, 2008.

[62] A. Cameron and P. Trivedi, *Regression Analysis of Count Data*, vol. 30. Cambridge Univ Pr, 1998.

[63] P. McCullagh and J. Nelder, *Generalized Linear Models*. Chapman & Hall/CRC, 1989.

[64] C. Croarkin and P. Tobias, "NIST/SEMATECH e-handbook of statistical methods," *National Institute of Standards and Technology*, 2006.

[65] N. Malo *et al.*, "Statistical practice in high-throughput screening data analysis," *Nature Biotechnology*, vol. 24, pp. 167–175, 2006.

[66] F. Markowetz, "How to understand the cell by breaking it: network analysis of gene perturbation screens," *PLoS computational biology*, vol. 6, p. e1000655, 2010.

[67] M. Robinson and A. Oshlack, "A scaling normalization method for differential expression analysis of RNA-seq data," *Genome Biol.*, vol. 11, p. R25, 2010.

[68] C. Soneson and M. Delorenzi, "A comparison of methods for differential expression analysis of rna-seq data," *BMC Bioinformatics*, vol. 14, no. 1, p. 91, 2013.

[69] L. Wang *et al.*, "DEGseq: An R package for identifying differentially expressed genes from RNA-seq data," *Bioinformatics*, vol. 26, pp. 136–138, 2010.

[70] J. Lloyd-Smith, "Maximum likelihood estimation of the negative binomial dispersion parameter for highly overdispersed data, with applications to infectious diseases," *PLoS One*, vol. 2, p. e180, 2007.

[71] W. Piegorsch, "Maximum likelihood estimation for the negative binomial dispersion parameter," *Biometrics*, vol. 46, pp. 863–867, 1990.

[72] N. Toft *et al.*, "The Gamma-Poisson model as a statistical method to determine if micro-organisms are randomly distributed in a food matrix," *Food microbiology*, vol. 23, pp. 90–94, 2006.

[73] D. McCarthy *et al.*, "Differential expression analysis of multifactor RNA-seq experiments with respect to biological variation," *Nucleic Acids Res.*, vol. 40, pp. 4288–4297, 2012.

[74] K. Bowman, "Extended moment series and the parameters of the negative binomial distribution," *Biometrics*, vol. 40, pp. 249–252, 1984.

[75] S. Clark and J. Perry, "Estimation of the negative binomial parameter κ by maximum quasi-likelihood," *Biometrics*, vol. 45, pp. 309–316, 1989.

[76] L. Willson *et al.*, "Multistage estimation compared with fixed-sample-size estimation of the negative binomial parameter k," *Biometrics*, pp. 109–117, 1984.

[77] C. Stein, "Inadmissibility of the usual estimator for the mean of a multivariate Normal distribution," in *Proceedings of the Third Berkeley symposium on mathematical statistics and probability*, vol. 1, pp. 197–206, 1956.

[78] W. James and C. Stein, "Estimation with quadratic loss," in *Proceedings of the fourth Berkeley symposium on mathematical statistics and probability held at the Statistical Laboratory, University of California, June 20-July 30, 1960*, p. 361, Univ of California Press, 1961.

[79] E. Lehmann and G. Casella, *Theory of Point Estimation*. Springer Verlag, 1998.

[80] J. Richards, "An Introduction to James-Stein estimation," 2009.

[81] B. Hansen, "Generalized shrinkage estimators," *Manuscript, University of Wisconsin*, 2008.

[82] G. Smyth, "Linear models and empirical Bayes methods for assessing differential expression in microarray experiments," *Statistical Applications in Genetics and Mol. Biol.*, vol. 3, p. 3, 2004.

[83] G. Smyth, "Limma: Linear models for microarray data," *Bioinformatics and Computational Biology Solutions Using R and Bioconductor*, pp. 397–420, 2005.

[84] Y. Benjamini and Y. Hochberg, "Controlling the False Discovery Rate: A practical and powerful approach to multiple testing," *JRSS(B)*, vol. 57, pp. 289–300, 1995.

[85] P. Hammer *et al.*, "mRNA-seq with agnostic splice site discovery for nervous system transcriptomics tested in chronic pain," *Genome Res.*, vol. 20, pp. 847–860, 2010.

[86] B. Tuch *et al.*, "Tumor transcriptome sequencing reveals allelic expression imbalances associated with copy number alterations," *PloS One*, vol. 5, p. e9317, 2010.

[87] T. M. Loughin, "A systematic comparison of methods for combining p-values from independent tests," *Computational Statistics and Data Analysis*, vol. 47, pp. 467–485, 2004.

[88] D. Bottomly *et al.*, "Evaluating Gene Expression in C57BL/6J and DBA/2J Mouse Striatum Using RNA-seq and Microarrays," *PloS one*, vol. 6, no. 3, p. e17820, 2011.

[89] A. Frazee *et al.*, "ReCount: A multi-experiment resource of analysis-ready RNA-seq gene count datasets," *BMC Bioinformatics*, vol. 12, p. 449, 2011.

[90] L. Shi *et al.*, "The MicroArray Quality Control (MAQC) project shows inter-and intraplatform reproducibility of gene expression measurements," *Nature Biotechnology*, vol. 24, pp. 1151–1161, 2006.

[91] B. Langmead *et al.*, "Ultrafast and memory-efficient alignment of short DNA sequences to the human genome," *Genome Biol.*, vol. 10, no. 3, p. R25, 2009.

[92] W. Zhining *et al.*, "Evaluation of gene expression data generated from expired Affymetrix GeneChip microarrays using MAQC reference RNA samples," *BMC Bioinformatics*, vol. 11, p. (Suppl 6):S10, 2010.

[93] E. Arikawa *et al.*, "Cross-platform comparison of SYBR Green real-time PCR with TaqMan PCR, microarrays and other gene expression measurement technologies evaluated in the MicroArray Quality Control (MAQC) study," *BMC Genomics*, vol. 9, p. 328, 2008.

[94] J. Bullard *et al.*, "Evaluation of statistical methods for normalization and differential expression in mRNA-seq experiments," *BMC Bioinformatics*, vol. 11, p. 94, 2010.

[95] T. Patterson *et al.*, "Performance comparison of one-color and two-color platforms within the MicroArray Quality Control (MAQC) project," *Nature Biotechnology*, vol. 24, pp. 1140–1150, 2006.

[96] M. Griffith *et al.*, "Alternative expression analysis by RNA sequencing," *Nature Methods*, vol. 7, pp. 843–847, 2010.

[97] A. Brooks *et al.*, "Conservation of an RNA regulatory map between Drosophila and mammals," *Genome Res.*, vol. 21, pp. 193–202, 2011.

[98] M. Sultan *et al.*, "A global view of gene activity and alternative splicing by deep sequencing of the human transcriptome," *Science*, vol. 321, p. 956, 2008.

[99] P. A. Swamy, "Efficient inference in a random coefficient regression model," *Econometrica: Journal of the Econometric Society*, pp. 311–323, 1970.

[100] M. Gumpertz and S. G. Pantula, "A simple approach to inference in random coefficient models," *The American Statistician*, vol. 43, no. 4, pp. 203–210, 1989.

[101] D. Bates, M. Maechler, and B. Bolker, "lme4: Linear mixed-effects models using s4 classes," 2012.

[102] C.-Y. Chang, P. Picotti, R. Hüttenhain, V. Heinzelmann-Schwarz, M. Jovanovic, R. Aebersold, and O. Vitek, "Protein significance analysis in selected reaction monitoring (srm) measurements," *Molecular & Cellular Proteomics*, vol. 11, no. 4, 2012.

[103] L. Reiter, O. Rinner, P. Picotti, R. Hüttenhain, M. Beck, M.-Y. Brusniak, M. O. Hengartner, and R. Aebersold, "mprophet: automated data processing and statistical validation for large-scale srm experiments," *Nature methods*, vol. 8, no. 5, pp. 430–435, 2011.

VITA

VITA

Danni Yu was born in Beijing, China. She received a B.S. in Economics from Capital University of Economics and Business, China in 2000. She received M.S. in Mathematical Statistics from Purdue University in 2008.